

# **Introduction à l'analyse biographique des durées**

**Support de formation 2024**

**Marc Thévenin**

2024-09-25

# Table des matières

<b>1</b>	<b>Présentation - Bibliographie - Outils</b>	<b>5</b>
1.1	Mises à jour 2024 . . . . .	5
	Le support . . . . .	5
	Bibliographie . . . . .	6
	Outils . . . . .	7
<b>I</b>	<b>Introduction</b>	<b>8</b>
<b>2</b>	<b>L'analyse biographique des durées</b>	<b>9</b>
2.1	Questions . . . . .	9
2.2	Terminologies . . . . .	9
2.3	Exemples d'analyse . . . . .	9
2.4	Elements nécessaire à l'analyse . . . . .	10
<b>II</b>	<b>Données et théorie</b>	<b>11</b>
<b>3</b>	<b>Les Données</b>	<b>12</b>
3.1	Données prospectives et rétrospectives . . . . .	12
3.1.1	Les données prospectives . . . . .	12
3.1.2	Les données rétrospectives . . . . .	13
3.2	Grille AGEVEN . . . . .	13
3.3	Enregistrement des données . . . . .	14
3.3.1	Large [Format individu] . . . . .	14
3.3.2	Semi-long [Format individu-événements] . . . . .	14
3.3.3	Long [Format individu-périodes] . . . . .	15
3.4	Exemples de mise à disposition . . . . .	15
3.4.1	Enquête biographie et entourage (Ined) . . . . .	15
3.4.2	Enquête MAFE (Ined) . . . . .	17
<b>4</b>	<b>La théorie</b>	<b>18</b>
4.1	Temps et durée . . . . .	18
4.2	Le Risk Set . . . . .	19
4.3	La Censure . . . . .	19
4.3.1	Censure à droite . . . . .	19
4.3.2	Censure à gauche, troncature et censure par intervalle . . . . .	21
4.4	Les grandeurs . . . . .	23
4.4.1	Les grandeurs utilisées . . . . .	23

4.4.2	La fonction de Survie $S(t)$ . . . . .	23
4.4.3	La fonction de répartition $F(t)$ . . . . .	24
4.4.4	La fonction de densité $f(t)$ . . . . .	24
4.5	Le risque instantané $h(t)$ . . . . .	24
4.5.1	Le risque cumulé $H(t)$ . . . . .	26
4.6	Remarques complémentaires . . . . .	28
4.6.1	Formes typiques de la fonction de survie . . . . .	28
4.6.2	Absence de censures à droites . . . . .	29
4.6.3	Utilisation des pondérations dans un schema retrospectif avec des biographies longues . . . . .	30
<b>III</b>	<b>Méthodes non paramétrique</b>	<b>32</b>
<b>5</b>	<b>Estimations des fonctions de survie</b>	<b>33</b>
5.1	Les fonctions de survie/séjour . . . . .	33
5.1.1	Les variables d'analyse . . . . .	33
5.1.2	Calcul de la fonction de survie . . . . .	34
5.2	La méthode actuarielle . . . . .	35
5.2.1	Estimation . . . . .	35
5.2.1.1	R . . . . .	36
5.2.1.2	Stata . . . . .	36
5.2.1.3	Python . . . . .	36
5.2.1.4	Sas <b>†</b> . . . . .	36
5.2.2	Application . . . . .	36
5.3	La méthode de Kaplan-Meier . . . . .	38
5.3.1	Estimation . . . . .	38
5.3.1.1	R . . . . .	39
5.3.1.2	Stata . . . . .	39
5.3.1.3	Python . . . . .	39
5.3.1.4	SAS <b>†</b> . . . . .	39
5.3.2	Application . . . . .	40
5.3.3	Quantités associées à l'estimateur Kaplan-Meier.. . . .	43
<b>6</b>	<b>Tests de comparaison</b>	<b>45</b>
6.1	Tests du log-rank . . . . .	45
6.1.1	Principe de calcul de la statistique de test . . . . .	45
6.1.2	Les principaux tests log-rank . . . . .	46
6.1.2.1	R . . . . .	47
6.1.2.2	Stata . . . . .	47
6.1.2.3	Python . . . . .	47
6.1.2.4	Sas <b>†</b> . . . . .	47
6.1.3	Application . . . . .	47
6.2	Comparaison des RMST . . . . .	48
6.2.0.1	R . . . . .	49
6.2.0.2	Stata . . . . .	49
6.2.0.3	Python . . . . .	49

6.2.0.4	SAS †	49
<b>IV</b>	<b>Modèles à risques proportionnels</b>	<b>52</b>
<b>7</b>	<b>Introduction aux modèles</b>	<b>53</b>
7.1	Proportionnalité des risques	53
7.2	Les modèles	54
<b>8</b>	<b>Le modèle de Cox</b>	<b>56</b>
8.1	Le modèle	56
8.1.1	La vraisemblance partielle et estimation des paramètres	56
8.1.2	Estimation des paramètres	58
8.1.3	Lecture des résultats	59
8.1.4	R	60
8.1.4.1	Stata	60
8.1.4.2	Python	60
8.1.4.3	SAS †	60
8.2	Analyse de la constance des rapports de risque	60
8.2.1	Test de Grambsch-Therneau sur les résidus de Schoenfeld	61
8.2.1.1	R	64
8.2.1.2	Stata	64
8.2.1.3	Python	64
8.2.1.4	SAS †	64
8.2.2	Intéraction avec la durée	64
8.2.3	Que faire ?	67
<b>9</b>	<b>Modèle à durée discrète</b>	<b>69</b>
9.1	Organisation des données	70
9.2	Ajustement de la durée	71
9.2.1	Ajustement avec une fonction quantitative de la durée	71
9.2.2	Ajustement discret	73
9.3	Proportionnalité des risques	74
<b>10</b>	<b>Variables dynamiques</b>	<b>76</b>
10.1	Facteur dynamique traitée de manière fixe	76
10.2	Estimation avec une variable dynamique	77
10.2.1	Modèle de Cox	77
10.2.1.1	R - Stata, Python	78
10.2.1.2	Sas †	78
10.2.2	Modèle à temps discret	79
10.3	Précautions à prendre	79

<b>V</b>	<b>Compléments</b>	<b>80</b>
<b>11</b>	<b>Éléments de mise en forme des données</b>	<b>81</b>
11.1	Calcul des variables d'analyses . . . . .	82
11.2	Appariement de modules biographiques . . . . .	88
11.2.1	Mise en forme d'une base . . . . .	88
11.2.2	Fusion des informations biographiques . . . . .	92
11.2.2.1	Fusion avec l'ensemble des périodes observables . . . . .	92
11.2.2.2	Fusion avec une autre base biographique . . . . .	95
11.3	Sélection d'un type de séquence et mise en forme pour l'analyse . . . . .	98
11.4	Durée jusqu'à la première séquence . . . . .	98
11.5	Durée de séjour dans la séquence d'intérêt et variables d'analyse . . . . .	102
<b>12</b>	<b>Risques concurrents</b>	<b>107</b>
12.1	Problématique . . . . .	107
12.2	Risques <i>cause-specific</i> et biais sur les estimateurs KM . . . . .	107
12.3	<b>Estimations en présence de risques concurrents (CIF)</b> . . . . .	109
12.3.1	Estimation non paramétrique . . . . .	109
12.3.2	Comparaison des CIF . . . . .	113
12.4	<b>Modèles</b> . . . . .	114
12.4.1	Modèles Semi paramétriques . . . . .	114
12.4.2	Modèle à temps discret . . . . .	114
<b>13</b>	<b>Modèles paramétriques</b>	<b>116</b>
13.1	Principes . . . . .	116
13.2	Hypothèse AFT: Accelerated Failure Time . . . . .	116
13.3	Principe de construction des modèles AFT . . . . .	117
13.4	Quelques modèles paramétriques usuels . . . . .	117
13.5	Exemple avec le modèle de Weibull . . . . .	119
13.6	Le modèle de Parmar-Royston . . . . .	119
<b>14</b>	<b>Annexes</b>	<b>121</b>
14.1	Tests Grambsch-Therneau OLS sur les résidus de Schoenfeld . . . . .	121
14.2	Fragilité et immunité . . . . .	124
14.2.1	Fragilité (Frailty) . . . . .	124
14.2.2	Immunité (Cure fraction) . . . . .	124
<b>VI</b>	<b>Programmation</b>	<b>126</b>
<b>15</b>	<b>R</b>	<b>127</b>
15.1	Packages et fonctions . . . . .	127
15.2	Analyse Non paramétrique . . . . .	128
15.2.1	Méthode actuarielle . . . . .	129
15.2.2	Méthode Kaplan-Meier . . . . .	130
15.2.3	Comparaison des S(t) méthode KM . . . . .	134

15.3	Modèle de Cox . . . . .	136
15.3.1	Estimation du modèle . . . . .	137
15.3.2	Hypothèse PH . . . . .	138
15.3.2.1	Test Grambsch-Therneau . . . . .	138
15.3.2.2	Introduction d'une interaction . . . . .	140
15.3.3	Introduction d'une variable dynamique (binaire) . . . . .	141
15.4	Modèles à durée discrète . . . . .	144
15.4.1	$f(t)$ quantitative . . . . .	146
15.4.2	$f(t)$ en indicatrices . . . . .	147
15.5	Modèles paramétriques usuels . . . . .	148
15.6	Risques concurrents . . . . .	150
15.6.0.1	Incidences cumulées . . . . .	151
15.6.0.2	Modèles . . . . .	154

# Liste des Figures

3.1	Biographie et entourage: base caractéristiques individuelles . . . . .	16
3.2	Biographie et entourage: base biographique logements . . . . .	16
3.3	MAFE: base caractéristiques individuelles . . . . .	17
3.4	MAFE: base biographique logement . . . . .	17
4.1	Schéma évènement/censure en temps calendaire . . . . .	21
4.2	Schéma évènement/censure sous forme de durée . . . . .	21
4.3	Grandeurs de la loi exponentielle avec $h(t)=0.04$ . . . . .	27
4.4	Fonction de survie: 3 situation typiques... ou pas . . . . .	28
4.5	Fonction de survie et modification de la métrique temporelle . . . . .	29
5.1	Courbe de survie: estimation méthode actuarielle . . . . .	38
5.2	Courbe de survie: estimation méthode actuarielle . . . . .	42
5.3	Courbe de survie: estimation méthode actuarielle + CI . . . . .	42
5.4	Risque cumulé: estimateur Nelson-Aalen . . . . .	43
5.5	Risque instantané: estimateur du Kernel . . . . .	44
6.1	Comparaison des Rmst à chaque jour où au moins un décès est observé . . . . .	51
7.1	L'hypothèse de proportionalite des risques . . . . .	54
9.1	Probabilité de décéder avec 3 ajustements de la durée . . . . .	72
9.2	Probabilité de décéder après correction de la non proportionnalité pour la variable surgery . . . . .	75
12.1	Fonction de répartition avec une cause concurrente traitée comme une censure à droite . . . . .	109
12.2	Risques concurrent: estimation de la CIF . . . . .	110
15.1	$S(t)$ méthode actuarielle avec <code>discSurv</code> (1) . . . . .	129
15.2	Méthode actuarielle avec <code>discSurv</code> (2) . . . . .	130

# Liste des Tables

5.2	Quantiles de la fonction de séjour type actuarielle - Bornes Sas . . . . .	37
5.3	Quantiles de la fonction de séjour type Kaplan-Meier . . . . .	42
6.1	Résultats des tests du logrank . . . . .	48
6.2	Estimation des Rmst pour la variable surgery . . . . .	50
6.3	Différences entre Rmst pour la variable surgery . . . . .	50
8.1	Cox: log Hazard Ratio (Risks Ratio) . . . . .	59
8.2	Cox: Hazard Ratio (Risks Ratio) . . . . .	59
8.3	Test OLS Grambsch-Therneau avec $g(t) = t$ . . . . .	63
8.4	Test Grambsch-Therneau avec $g(t) = 1 - S(t)$ . . . . .	63
8.5	Base spittées sur les intervals d'évènement . . . . .	65
8.6	Modèle de Cox avec une interaction entre une fonction de la durée et la variable *surgery . . . . .	65
9.1	Durée discrète: données en format d'origine . . . . .	71
9.2	Durée discrète: données en format long . . . . .	71
9.4	Modèle logistique à durée discrète ( $f(t)$ continue) . . . . .	72
9.5	Modèle de Cox . . . . .	73
9.6	Modèle logistique à durée discrète ( $f(t)$ indicatrices) . . . . .	74
9.8	Modèle logistique à durée discrète avec correction de la non proportionnalité . . . . .	75
10.1	Modèle de cox avec une variable dynamique (binaire) traitée de manière fixe (estimation biaisée . . . . .	76
10.2	Mapping de la base avec une variable dynamique binaire traitée de manière fixe . . . . .	77
10.3	Mapping correct de la base avec une variable dynamique binaire . . . . .	77
10.4	Modèle de Cox avec une variable dynamique binaire . . . . .	78
10.5	Modèle logistique à durée discrète avec variable dynamique binaire . . . . .	79
12.1	Test de Gray pour la variable surgery . . . . .	113
12.2	Test de Pepe-Mori pour la variable surgery . . . . .	113
12.3	Modèle logistique multinomial avec risques concurrent . . . . .	115
13.1	Modèle de Weibull . . . . .	118
13.2	Modèle de Parmar-Roytston . . . . .	120



# 1 Présentation - Bibliographie - Outils

## 1.1 Mises à jour 2024

### Test de Grambsch Therneau l'hypothèse de proportionnalité dans un modèle de Cox: version exacte versus approximation OLS

Une contribution majeure visant à expliquer les écarts observés entre les deux versions du test a été réalisée fin 2023. [Lien](#) Elle repose sur la présence de corrélations même limitées entre les covariables du modèle, situation classique en sciences sociales. La conclusion de ce travail penche clairement vers une utilisation du test ols (moindre carrés ordinaires) dans le domaine des sciences sociales.

Dans ce support, j'ai donc résumé le travail de S.Metzger dans la section dédiée au test de proportionnalité avec un modèle de Cox. Comme cela impacte le travail sous R j'ai également repris la section relative à la programmation, bien que j'avais déjà donné le moyen de récupérer et d'utiliser le test ols depuis mai 2022.

### Calcul des Rmst avec Python

Ajouté au package `lifelines` récemment, des éléments de programmation ont été ajoutés dans la section Python. La syntaxe n'est pas très conviviale, mais c'est déjà ça. Le test de comparaison des Rmst n'y pas implémenté, ce qui est dommage vu sa simplicité. Plus généralement, aucune inférence de l'estimateur ne semble être présente dans la fonction.

### Quelques ajouts sur l'épineuse question des pondérations en analyse de survie

Cela fait suite à des questionnements qui m'ont été adressés par le groupe d'exploitation de l'enquête [Envie](#). Mon court argumentaire qui reposait sur l'utilisation des pondérations avec des biographies longues a été, il faut l'avouer, un peu mis à mal, la plage d'âges étant ici de 18 à 29 ans. Cela m'a permis de faire un tour d'horizon plus complet de la problématique, et qui continue bien évidemment à me renvoyer dans le domaine de la médecine.

## Le support


Ce document est utilisé comme support de formation, principalement pour celle dispensée à l'Ined, celle effectuée dans le cadre d'HED, ainsi que le cours de master 2 de démographie de l'Université de Strasbourg. En terme de contenu il reste classique, il s'agit d'une introduction, même si certains *apports* méthodologiques plus ou moins récents sont introduits comme l'estimation des **RMST**<sup>1</sup>. Une méthode modélisation reposant sur des pseudo-observations<sup>2</sup> est en cours d'évaluation, et pourrait être

---

<sup>1</sup>Restricted Mean of Survival Time

<sup>2</sup>Résidus du Jackknife

introduite dans le support fin 2024. Mais... Conceptuellement très intéressante, son application dans les sciences sociales risque de régulièrement buter sur une hypothèse restrictive, à savoir la non corrélation entre observations censurées et covariables. Elle sera souvent remise en cause lorsqu'on introduira des dimensions explicatives tels que l'âge ou la génération en début d'exposition, soit des éléments très communs dans les analyses.

Sur la forme, le support a été passé en 2023 en format ouvrage (sans en avoir l'ambition il est bon de le préciser) et une version pdf peut-être directement téléchargée. elle est identique à la version html, sauf pour le chapitre programmation ou seulement les éléments relatifs à R sont présents. Les autres outils sont **Stata**, **Python** (Lifelines et Statmodel), et **Sas**  sont propres à la version html.

Enfin un petit mot sur l'application présente dans le support. Issu du champ de la médecine, soit l'analyse de la survie de personnes souffrant d'une insuffisance cardiaque, elle peut décevoir vu son éloignement avec des problématiques issues des sciences sociales<sup>3</sup>. Cependant, cette base d'analyse permet de couvrir avec peu d'informations, la quasi totalité des points traités dans ce support. Elle repose sur un nombre d'observations très limité (103), collectées il y a quand même très longtemps (fin 60 début 70), et avec seulement 4 covariables. Il y a sans aucun doute des biais un peu partout, en partie expliquée par non prise en compte de variables de contrôle, pure ou de confusion. J'admets qu'il serait préférable de trouver un jour autre chose, ou de donner par exemple en annexe, quelques exemples d'applications plus proches des sciences sociales. Néanmoins pour les personnes participant à la formation, les jeux de données utilisés pour les travaux pratiques sont bien issus des sciences sociales.

Il ne s'agit pas d'un support validé institutionnellement. J'en assume donc totalement seul le contenu. J'en profite également pour remercier, quelques soient leur statut et leur fonction, l'ensemble des personnes ayant eu recours à mon assistance sur ce champ d'analyse très riche, ainsi que l'ensemble les participant.e.s aux formations et cours. Par leurs remarques, les problématiques traitées, ils me permettent de réviser et mettre à jour annuellement ce document. Enfin, je remercie plus particulièrement [Eva Lelièvre](#) et [Arnaud Bringé](#).

## Bibliographie

Les éléments bibliographiques qui figurent ci-dessous proviennent du champ des sciences sociales. Elle est volontairement courte, mais efficace. Quelle que soit la langue, le nombre de cours ou support sont très nombreux en médecine. On trouve également de trop nombreux tutoriels généraliste à dominante *mise en pratique avec R*, dont je ne conseille pas forcément l'utilisation.

### Accès en ligne

- **Cours Gilbert Colletaz** (Université d'Orléans - Master d'économétrie).
  - Applications uniquement avec Sas. Suite à la retraite de son auteur, le cours n'est plus mis à jour.
  - Dernière version 2020: [lien](#).
- **Document de travail de Simon Quantin** (Insee).

---

<sup>3</sup>Le document de travail de l'Insee indexé dans la bibliographie ne fait pas mieux, l'application traite du diabète

- Couvre l’ensemble des techniques de base d’analyse des durées en durée dite continue. Il propose surement la meilleure introduction en langue française à la problématique de la *fragilité*, qui sera ici seulement présenté trop brièvement.
- Application en R seulement (Attention au passage de la v3 du package **survival** pour la question du test de proportionnalité de Grambsch-Therneau).
- 2019 - pas de mise à jour: [lien](#).
- **Les notes de cours de German Rodriguez** (en)
  - Démographie à l’université de Princeton.
  - Les dernières mises à jour doivent dater de 2017-2018: [lien](#)

Ouvrage de référence en démographie:

- ***L’analyse démographique des biographies*** de *Daniel Courgeau* et *Eva Lelièvre* (Edition de l’Ined - 1989). Malheureusement cet ouvrage ne dispose pas de version epub ou pdf disponible en ligne <sup>4</sup>.

## Outils

- Support réalisé sous [Rstudio](#) avec l’outil d’édition [Quarto](#)
- Langages utilisés pour la partie programmation:
  - R
  - [Stata v18](#)
  - [Python](#)
  - [+ Sas v9](#)




---

<sup>4</sup>Pour les résident.e.s du campus Condorcet, l’ouvrage est disponible au GED [\[lien\]](#)

**partie I**

**Introduction**

## 2 L'analyse biographique des durées

### 2.1 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/changer au cours de la période d’observation: variables dynamiques (**TVC**: *Time Varying Covariate*)

### 2.2 Terminologies

Français	Anglais
Analyse des durées	Duration analysis
Analyse de survie/séjour	Survival analysis
Analyse de fiabilité	Failure time data analysis
Analyse des transitions	Event-history analysis

### 2.3 Exemples d'analyse

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

## 2.4 Elements nécessaire à l'analyse

### 1. Un processus temporel

- Une échelle de mesure ou métrique temporelle: minutes, heures, jours, mois, années....
- Une origine **commune** définissant un évènement de départ <sup>1</sup>: naissance, mariage si on analyse la séparation, ....
- Une définition précise de l'évènement d'étude.
- Une durée entre le début et la fin de la période d'observation, si nécessaire avec la fin de la période d'exposition au risque. Cette sortie d'exposition devrait être également calculée à l'aide des informations de datation, ce qui n'est pas toujours le cas.

### 2. Une population soumise au risque de connaître l'évènement (Risk Set)

### 3. *Des variables explicatives* ou *covariables*

- Fixes: sexe/genre, génération, niveau de diplôme le plus élevé,..... Ce sont généralement des caractéristiques en début d'exposition.
- Dynamiques (TVC: *Time varying covariates*):
  - Mesurées à tout moment entre le début et la sortie de l'observation: statut matrimonial, taille du ménage, statut d'activité...
  - Pour les modèles, et à l'exception du semi-paramétrique de Cox, la durée ou une transformation de celle-ci, est **une variable dynamique introduite comme variable indépendante pour assurer le bon ajustement des données**. L'introduction directe d'une fonction de la durée comme variable dépendante seule ne peut se faire qu'en absence d'observation censurée, en particulier à droite. Il s'agit là d'une caractéristique propre aux modèles pleinement paramétriques: usuels basés sur une loi de distribution des évènements dans le temps (Weillbul, Gompertz...) ou les modèles dits à durée discrète (logit, probit...).

---

<sup>1</sup>Attention, dans le cadre des données prospectives ou de suivi, cela ne peut pas être le moment de l'inclusion à la base données

# **partie II**

## **Données et théorie**

## 3 Les Données

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

### 3.1 Données prospectives et rétrospectives

#### 3.1.1 Les données prospectives

- Individus suivis à des dates successives. On parle souvent de données de stock mises à jour à intervalles de temps plus ou moins réguliers.
- Instrument de mesure identique à chaque vague (si possible).
- **Avantages:**
- Qualité des données et techniquement l'absence de biais de mémoire<sup>1</sup>.
- Si le suivi est pérenne une même analyse peut être répliquée à intervalles réguliers.
- Inconvénients:
- Délais pour les exploiter dans une analyse. Mais à minima deux points d'observations permettent déjà sur une exposition certes très courte, de présenter quelques résultats.
- Mêmes hypothèses entre deux passages pas forcément respectées
- Attrition, censure ou troncature à gauche liés aux âges d'inclusion. C'est sans aucun doute le plus gros problème, et ces phénomènes demande une vigilance extrême. Sans connaissance des principes de base en analyse des durées ou de survie, on peut être amené à réaliser, évaluer ou lire des études que l'on pourrait qualifier ici d'*analyse de survie Canada dry*.

A noter l'exploitation croissante des données administratives qui peuvent regorger d'informations biographiques. Déjà disponibles, le problème du coût de collecte est contourné<sup>2</sup>. Ce type de données comprend par exemple les informations issues des fichiers des Ressources Humaines des entreprises, qui ont été exploitées à l'Ined par exemple dans le cadre du projet « worklife » (<https://worklife.site.ined.fr/>). Une des sources de plus en plus utilisée en France est maintenant l'EDP<sup>3</sup>. Elles engendrent en revanche des

---

<sup>1</sup>Cet avantage peut se trouver contrebalancé par des phénomènes de censure par intervalles, donc de *trous* tout aussi problématiques que ceux liés à la mémoire

<sup>2</sup>Je ne suis pas forcément à l'aise avec cet argument souvent avancé. La maintenance et l'alimentation de ce type de données peut être également coûteuse, ne se faisant pas par magie, comme pour l'EDP de l'Insee

<sup>3</sup>Echantillon Démographique Permanent. Un bon exemple de données administrative dont le coût de production est loin d'être négligeable



questionnements techniques liés à l'inférence (on ne travaille directement pas sur des échantillons), et à une présence potentiellement massive de problèmes de censures à gauche ou par intervalles, ou de troncature à gauche<sup>4</sup>.

### 3.1.2 Les données rétrospectives

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

Les deux types de recueil peuvent être mixés avec des enquêtes à passages répétés comprenant des informations rétrospectives entre 2 vagues. Par exemple la cohorte [Elfe](#) de l'Ined-Inserm ou la [Millenium-Cohort-Study](#) en Grande Bretagne.

## 3.2 Grille AGEVEN

Pour recueillir des informations biographiques rétrospectives, 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 dans l'article de G.Vivier, [page 121](#)

---

<sup>4</sup>Se reporter par exemple à la présentation du très rigoureux guide de l'utilisateur de l'EDP: *Les informations disponibles : des informations à géométrie variable et à trous*

## 3.3 Enregistrement des données

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

### 3.3.1 Large [Format individu]

Une ligne par individu qui renseigne tous les événements liés à un domaine: les datations et 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'événements. 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.

Remarque: pour des enquêtes non biographiques, mais avec quelques éléments de datation qui s'intéresse par exemple à la date d'une *première fois* [J'écris cela en 2024, donc je pense par exemple à l'exploitation de l'enquête [Envie](#).

### 3.3.2 Semi-long [Format individu-événements]

C'est le format le plus courant de mise à disposition des enquêtes biographiques. Si les transitions sont de type continu, par exemple le lieu de résidence (on habite toujours quelque part), la date de fin de la séquence correspond à la date de début de la séquence suivante. Les dates de fin ne sont pas forcément renseignées sur une ligne pour des trajectoires continues, l'information peut être donnée sur la ligne suivante avec la date de début.

Pour la séquence qui se déroule au moment de l'enquête, la date de fin est souvent une valeur manquante, une fin de séquence pouvant se produire juste avant l'enquête au cours d'une même année. Il est également possible d'avoir une information qui ne s'est pas encore produite au moment de l'enquête, mais qui aura lieu peu de temps après (personne enceinte, donc une naissance probable la même année).

Exemple précédent (trajectoires discontinues):

id	debut	fin	cause	Numero séquence
A	1979	1982	décès conjoint	1
A	1985	.	.	2

id	debut	fin	cause	Numero séquence
B	1983	1984	Séparation	1

### 3.3.3 Long [Format individu-périodes]

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

Exemple précédent:

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

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

## 3.4 Exemples de mise à disposition

Deux gros classiques d'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.

### 3.4.1 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/)

Figure 3.1: Biographie et entourage: base 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	Nationalité actuelle en clair
1	101	ANDREE		2 06/19/1938	93	LIVRY-GARGAN		46	FRANCAISE
2	102	JEANINE		2 06/11/1934	37	TOURS		261	FRANCAISE
3	103	MANUEL		1 08/20/1942	99	NR	PORTUGAL	99139	PORTUGAISE
4	104	LEON		1 01/13/1933	93	BONDY		10	FRANCAISE
5	105	FRANCOIS		1 12/27/1932	99	ALGER	ALGERIE	99352	FRANCAISE
6	106	EVELYNE		2 11/21/1950	99	NR	ALGERIE	99352	FRANCAISE
7	107	MICHEL		1 05/23/1949	75	PARIS-20E__ARRONDISSEMENT		120	FRANCAISE
8	108	JEANNINE		2 05/21/1948	94	PERREUX-SUR-MARNE		58	FRANCAISE
9	109	BEATRICE		2 06/09/1949	59	LOUVROIL		365	FRANCAISE
10	110	THANH CUA		1 03/16/1941	99	TRAVINH	VIET NAM	99243	FRANCAISE
11	111	MAXIME		1 07/31/1950	77	LAGNY-SUR-MARNE		243	FRANCAISE
12	112	JACQUELINE		2 09/25/1934	54	SAINT-MAX		482	FRANCAISE
13	113	YVETTE		2 09/09/1937	19	CORNIL		61	FRANCAISE
14	114	ZOFIA		2 06/11/1935	99	EMILOWNA	POLOGNE	99122	POLONAISE
15	115	ANTONIO		1 09/19/1932	99	SEVILLE	ESPAGNE	99134	ESPAGNOL
16	116	JEAN PIERRE		1 04/18/1930	75	PARIS-12E__ARRONDISSEMENT		112	FRANCAISE
17	117	JOSETTE		2 04/20/1939	75	PARIS- 6E__ARRONDISSEMENT		106	FRANCAISE
18	118	RADA		2 12/18/1945	99	ZAGREB	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	ESQUERDES		309	FRANCAISE
23	123	DANIEL		1 06/12/1948	75	PARIS-14E__ARRONDISSEMENT		114	FRANCAISE
24	124	JEAN-CLAUDE		1 08/31/1936	92	NEUILLY-SUR-SEINE		51	FRANCAISE
25	125	GHISLAINE		2 01/20/1944	60	BRETEUIL		104	FRANCAISE
26	126	JOCELYNE		2 06/28/1949	28	BOULLAY-LES-DEUX- EGLISES		53	FRANCAISE
27	127	MARIE-JOSE		2 10/31/1949	76	MONT-SAINT-AIGNAN		451	FRANCAISE

Figure 3.2: Biographie et entourage: base biographique logements

VIEWTABLE: TMP1.tego									
	Identifiant questionnaire	prénom d ego	sexe d ego	Date de naissance	Département de naissance	Commune ou pays de naissance	Pays ou DOM-TOM de naissance	Numéro INSEE de la commune de naissance	Nationalité actuelle en clair
1	101	ANDREE		2 06/19/1938	93	LIVRY-GARGAN		46	FRANCAISE
2	102	JEANINE		2 06/11/1934	37	TOURS		261	FRANCAISE
3	103	MANUEL		1 08/20/1942	99	NR	PORTUGAL	99139	PORTUGAISE
4	104	LEON		1 01/13/1933	93	BONDY		10	FRANCAISE
5	105	FRANCOIS		1 12/27/1932	99	ALGER	ALGERIE	99352	FRANCAISE
6	106	EVELYNE		2 11/21/1950	99	NR	ALGERIE	99352	FRANCAISE
7	107	MICHEL		1 05/23/1949	75	PARIS-20E__ARRONDISSEMENT		120	FRANCAISE
8	108	JEANNINE		2 05/21/1948	94	PERREUX-SUR-MARNE		58	FRANCAISE
9	109	BEATRICE		2 06/09/1949	59	LOUVROIL		365	FRANCAISE
10	110	THANH CUA		1 03/16/1941	99	TRAVINH	VIET NAM	99243	FRANCAISE
11	111	MAXIME		1 07/31/1950	77	LAGNY-SUR-MARNE		243	FRANCAISE
12	112	JACQUELINE		2 09/25/1934	54	SAINT-MAX		482	FRANCAISE
13	113	YVETTE		2 09/09/1937	19	CORNIL		61	FRANCAISE
14	114	ZOFIA		2 06/11/1935	99	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	ESQUERDES		309	FRANCAISE
23	123	DANIEL		1 06/12/1948	75	PARIS-14E__ARRONDISSEMENT		114	FRANCAISE
24	124	JEAN-CLAUDE		1 08/31/1936	92	NEUILLY-SUR-SEINE		51	FRANCAISE
25	125	GHISLAINE		2 01/20/1944	60	BRETEUIL		104	FRANCAISE
26	126	JOCELYNE		2 06/28/1949	28	BOULLAY-LES-DEUX- EGLISES		53	FRANCAISE
27	127	MARIE-JOSE		2 10/31/1949	76	MONT-SAINT-AIGNAN		451	FRANCAISE

### 3.4.2 Enquête MAFE (Ined)

Figure 3.3: MAFE: base 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	Nationalité actuelle en clair
1	101	ANDREE		2 06/19/1938	93	LIVRY-GARGAN		46	FRANCAISE
2	102	JEANINE		2 06/11/1934	37	TOURS		261	FRANCAISE
3	103	MANUEL		1 08/20/1942	99	NR	PORTUGAL	99139	PORTUGAISE
4	104	LEON		1 01/13/1933	93	BONDY		10	FRANCAISE
5	105	FRANCOIS		1 12/27/1932	99	ALGER	ALGERIE	99352	FRANCAISE
6	106	EVELYNE		2 11/21/1950	99	NR	ALGERIE	99352	FRANCAISE
7	107	MICHEL		1 05/23/1949	75	PARIS-20E_ARRONDISSEMENT		120	FRANCAISE
8	108	JEANNINE		2 05/21/1948	94	PERREUX-SUR-MARNE		58	FRANCAISE
9	109	BEATRICE		2 06/09/1949	59	LOUVROIL		365	FRANCAISE
10	110	THANH CUA		1 03/16/1941	99	TRAVINH	VIET NAM	99243	FRANCAISE
11	111	MAXIME		1 07/31/1950	77	LAGNY-SUR-MARNE		243	FRANCAISE
12	112	JACQUELINE		2 09/25/1934	54	SAINT-MAX		482	FRANCAISE
13	113	YVETTE		2 09/09/1937	19	CORNIL		61	FRANCAISE
14	114	ZOFIA		2 06/11/1935	99	EMILOWNA	POLOGNE	99122	POLONAISE
15	115	ANTONIO		1 09/19/1932	99	SEVILLE	ESPAGNE	99134	ESPAGNOL
16	116	JEAN PIERRE		1 04/18/1930	75	PARIS-12E_ARRONDISSEMENT		112	FRANCAISE
17	117	JOSETTE		2 04/20/1939	75	PARIS-6E_ARRONDISSEMENT		106	FRANCAISE
18	118	RADA		2 12/18/1945	99	ZAGREB	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	ESQUERDES		309	FRANCAISE
23	123	DANIEL		1 06/12/1948	75	PARIS-14E_ARRONDISSEMENT		114	FRANCAISE
24	124	JEAN-CLAUDE		1 08/31/1936	92	NEUILLY-SUR-SEINE		51	FRANCAISE
25	125	GHISLAINE		2 01/20/1944	60	BRETEUIL		104	FRANCAISE
26	126	JOCELYNE		2 06/28/1949	28	BOULLAY-LES-DEUX- EGLISES		53	FRANCAISE
27	127	MARIE-JOSE		2 10/31/1949	76	MONT-SAINT-AIGNAN		451	FRANCAISE

Figure 3.4: MAFE: base biographique logement

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

Quelques éléments de manipulation de ce type de données biographiques sont présentés dans le chapitre compléments<sup>5</sup> [\[lien\]](#)

<sup>5</sup>Il s'agit d'un premier jet réalisé pour la version 2023 et qui ne peut pas viser l'exhaustivité

## 4 La théorie

L'analyse des durées peut être vue comme l'étude d'une variable aléatoire  $T$  qui décrit la durée d'attente jusqu'à l'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**, lorsque la **durée d'observation est plus courte que la durée d'exposition au risque**.

### 4.1 Temps et durée

Le temps est une dimension (la quatrième), la durée est sa mesure. La durée est tout simplement calculée par la différence, pour une échelle temporelle donnée, entre la fin et le début d'une période d'exposition ou d'observation.

On distingue généralement deux types de mesure de la durée : **continue** et **discrète/groupée**. Ces deux notions ne possèdent pas réellement de définition, la différence s'explique plutôt par la présence ou non de simultanéité dans l'occurrence des événements. Le temps est intrinsèquement continu car deux événements ne peuvent pas avoir lieu en « même temps ». C'est donc l'échelle temporelle choisie ou imposée par l'analyse et les données qui pourra rendre cette mesure continue ou discrète/groupée. Pour un physicien travaillant sur la théorie de la relativité avec des horloges atomiques, une minute (voire une seconde) est une mesure très groupée pour ne pas dire grossière du temps, alors que pour un géologue c'est une mesure continue. Pour ces deux disciplines, cette échelle de mesure n'est pas adaptée à leur domaine. Le choix de l'échelle temporelle doit être pertinent par rapport aux objectifs de l'analyse même si on dispose des informations très fines (dates de naissance exactes). Etudier la fécondité avec une métrique journalière qui n'aurait pas de sens.

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

#### ! Important

- **Durée continue** : absence (ou très peu) d'événements mesurés simultanément
- **Durée discrète/groupée** : présence constante et/ou en grand nombre d'événements simultanés

## 4.2 Le Risk Set

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

## 4.3 La Censure

! Important

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

### 4.3.1 Censure à droite

#### Définition

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

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

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

En pratique (important)

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

## Exemple

On effectue une enquête auprès de femmes : On souhaite mesurer l'âge à la naissance de leur premier enfant. Au moment de l'enquête, une femme est âgée de 29 ans n'a pas 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.

La question se posera différemment si la personne à 75 ans et n'a pas d'enfant, car elle sera sortie de l'exposition au risque.

## Hypothèse fondamentale

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

### *Problème posé par la censure informative*

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

- A l'Ined l'exploitation du registre des personnes atteintes de mucoviscidose (G.Belis) donne une autre illustration de ce phénomène. Chaque année un nombre significatif de personnes sortent du registre. On pas les résultats des examens annuels qu'ils doivent subir. Si certain.e.s perdu.e.s de vue s'expliquent par des déménagements, émigration ou par un simple problème d'enregistrement des informations (le médecin a oublié), on note qu'ils/elles sont nombreu.se.s à présenter une forme « légère » de la maladie. Cette information étant être donnée ici par la mutation du gène. Comme il n'est pas recommandé de supprimer ou de traiter ces observations comme des censures à droite non informatives, on peut les appréhender comme un risque concurrent au décès ou à tout autre évènement analysé à partir de ce registre (voir section dédiée).

Les graphiques suivant représentent, en temps calendaire et après sa transformation en durée, la logique des censures à droite. Le recueil des informations est ici de nature prospectives, et bien évidemment on suppose que le début de l'observation correspond à un début d'exposition cohérent avec l'analyse réalisée (année de diagnostic d'une maladie, début d'une séquence d'emploi, de lieu de résidence, de couple ou de célibat strict etc....).

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



Figure 4.1: Schéma évènement/censure en temps calendaire

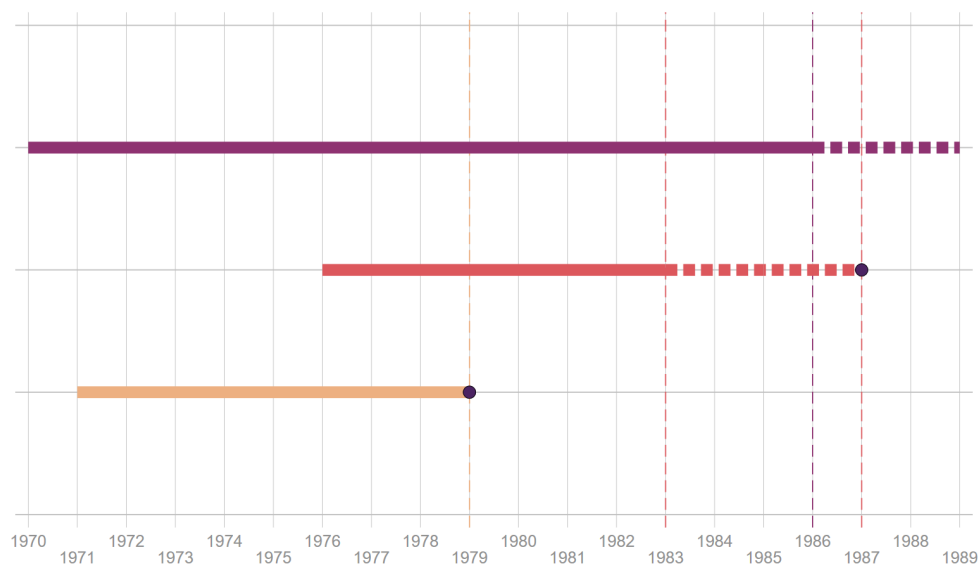
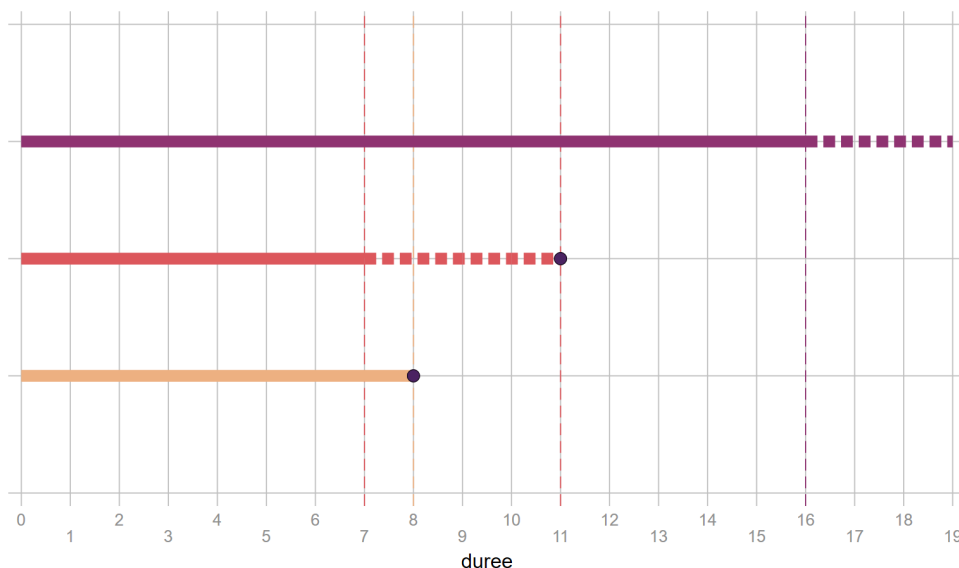


Figure 4.2: Schéma évènement/censure sous forme de durée



### 4.3.2 Censure à gauche, troncature et censure par intervalle

#### Censure à gauche

L'évènement a pu se produire avant le début période d'observation, mais on est pas en mesure de savoir s'il s'est produit, et si on sait qu'il s'est produit on est pas en mesure de savoir quand. Typique des données prospectives, de type registre, avec par exemple des âges à l'inclusion différenciés. La présence de ce type de censure ne permet de définir lors de la création de l'échantillon d'analyse des durées d'exposition cohérente pour l'ensemble de observations de départ. Même si elle ne sont pas traitée dans ce support, il existe quelques méthodes, en durée continue seulement, pour obtenir des résultats

en présence de ce type de censure, mais à la condition qu'elle ne soit pas trop nombreuses. On peut également filtrer l'échantillon en amont en conservant seulement celles dont le début d'exposition est clairement défini<sup>1</sup>.

### Censure par intervalle

Traditionnellement la censure par intervalle est définie par l'impossibilité de dater exactement la survenue d'un événement dans un intervalle de temps<sup>2</sup>. Dans ce sens, on pourrait donc affirmer qu'elle est une caractéristique propre aux temporalités groupées dites à durée discrète<sup>3</sup>. On peut sans problème généraliser ce phénomène de censure à l'occurrence: un événement peut se produire entre 2 temps d'observations sans qu'on puisse l'observer. Un exemple classique en criminologie est celle de la récidive d'un délit entre deux arrestations ou deux condamnations: on sait qu'une personne a récidivé en raison de son arrestation, mais on est pas en mesure de savoir s'elle a récidivé plus tôt....car pas vu pas pris. A noter également, toujours dans un recueil prospectif, qu'un phénomène de censure à droite lié aux *perdu.e de vue* peut se transformer en censure par intervalle lorsque la personne "réapparaît" et est de nouveau incluse à l'échantillon<sup>4</sup>.

### Troncature

Par l'exemple, on analyse la survie d'une population. Seule la survie des individus vivants à l'inclusion peut être analysée (*troncature à gauche*). On peut donc rencontrer un phénomène de sélection difficilement contrôlable.

Un exemple: dans l'enquête [Virage](#), on a des informations sur les tentatives de suicide. Seule l'analyse des tentatives de suicide n'ayant pas entraîné le décès peuvent être analysés.

### Durée d'observation supérieure à la durée d'exposition

A l'inverse de la censure, 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.

- 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.
- Situation traitée dans le TP de la formation: on analyse la durée de la première séquence d'emploi ou d'une suite de séquence d'emploi sans rupture (chômage, maladie, sortie du marché du travail etc...). Il conviendra pour les personnes qui n'ont pas connu de rupture (d'au moins un an par exemple) de faire sortir certaines personnes au moment de la retraite et non au moment de l'enquête, si elle sont déjà retraitées lors du recueil. Compte tenu des générations enquêtées, plutôt anciennes, on pourra considérer ces âges à la retraite comme des âges à censures à droites non informatives.

---

<sup>1</sup>Ce qui a été fait dans le projet worklife pour Air France

<sup>2</sup>Garder en mémoire que l'analyse des durées ou de survie a été très largement développé dans un cadre à durée continue

<sup>3</sup>Si on utilise une mesure de la durée sur l'âge, on ne sait pas si l'événement a eu lieu le lendemain de l'anniversaire ou la veille de l'anniversaire suivant

<sup>4</sup>Voir exemple plus haut sur le registre de la mucoviscidose

## 4.4 Les grandeurs

### 4.4.1 Les grandeurs utilisées

- La fonction de survie:  $S(t)$
- La fonction de répartition:  $F(t)$
- La fonction de densité:  $f(t)$
- Le risque *instantané*:  $h(t)$
- Le risque *instantané* cumulé:  $H(t)$

#### Remarques:

- Toutes ces grandeurs sont mathématiquement liées les unes par rapport aux autres. En connaître une permet d'obtenir les autres.
- Au niveau formel ***on se placera ici du point de vue où la durée mesurée est strictement continue***. Cela se traduit, entre autre, par l'absence d'événements dits "simultanés". En présence de durée discrète/groupée, il est à noter que les expressions se simplifient, en particulier pour la densité ou le risque dit *instantané*.
- Les expressions qui vont suivre ne sont pas des estimateurs, mais des grandeurs dont on précisera seulement les propriétés. Ce sont les techniques d'estimations, et la cohérence des données avec la théorie, qui devront respecter ces propriétés.

### 4.4.2 La fonction de Survie $S(t)$

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

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

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

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

Propriétés:

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

La fonction de survie est donc strictement non croissante.

### 4.4.3 La fonction de répartition $F(t)$

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

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

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

La fonction de survie et la fonction de répartition sont donc deux grandeurs strictement complémentaires, pour ne pas dire identique, et décrivent la même information.

Propriétés:

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

### 4.4.4 La fonction de densité $f(t)$

- Pour une valeur de  $t$  donnée, la fonction de densité de l'évènement donne la distribution des moments où les évènements ont eu lieu. Le numérateur donne classique, densité oblige, la probabilité de connaître l'évènement dans un petit intervalle de temps  $dt$ . Si  $dt$  est proche de 0 (temps continu) alors cette probabilité tend également vers 0. On norme donc cette probabilité par  $dt$ . Rappel: on est toujours ici dans la théorie.
- En temps continu, la fonction de densité est donnée par la dérivée de la fonction de répartition:  $f(t) = F'(t) = -S'(t)$ . On reste dans des relations statistiques élémentaires... mais cette densité n'est pas formellement une probabilité. Il s'agit plutôt d'un taux.

Formellement la fonction de densité  $f(t)$  s'écrit:

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

## 4.5 Le risque instantané $h(t)$

Concept fondamental de l'analyse des durées:

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

- $P(t \leq T < t + dt | T \geq t)$  donne la probabilité de survenue de l'évènement sur l'intervalle  $[t, t + dt[$  *conditionnellement à la survie au temps  $t$ .*
- En divisant par  $dt$ , La quantité obtenue donne alors un nombre moyen d'évènements que connaîtrait un individu durant un très court intervalle de temps.

- A priori cette quantité n'est pas une probabilité. C'est la nature de l'évènement, en particulier sa non récurrence, et la métrique temporelle choisie ou disponible qui peut la rendre assimilable à une probabilité. Tout comme la densité, on est plutôt dans la définition d'un *taux* (d'où l'expression ***hazard rate*** en anglais... traduction ??? taux de risque ???).

On peut également écrire:

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)} \quad 5$$

On remarque que cette fonction de risque (ou *hazard rate*) n'est pas une probabilité car  $\frac{f(t)}{S(t)}$  ne peut pas contraindre ici la valeur obtenue à ne pas être supérieure à 1.

### ! Grandeurs avec des durées discrètes/groupées

Les expressions de la densité et du risque dit instantané se simplifient:  $f(t) = P(t \leq T < t + dt)$  et  $h(t) = P(t \leq T < t + dt | T \geq t)$ . Ces deux grandeurs ne sont plus des taux tel qu'ils ont été définis précédemment, mais des probabilités. En durée discrète, on aurait donc  $dt = 1$

Néanmoins les relations entre les grandeurs, en particulier celle qui lie risque, densité et survie reste toujours valable. Ceci est fondamental, car elle permet de définir la fonction de vraisemblance sur laquelle repose le calcul de tous les estimateurs.

### Illustration du concept

Cette notion de *taux de risque* ou *hazard rate* non réductible à une probabilité, est à mon sens très bien illustré par G.Colletaz dans ces notes de cours (pages 11-12). Dans ce qui suit, j'en fais un quasi copier-coller. En se positionnant sur des échelles temporelles plus facilement saisissables pour la description de phénomènes socio-démographiques (on supprime la limite des expressions données pour  $f(t)$  et  $h(t)$ ).

1. On s'intéresse au risque d'attraper un rhume durant les mois d'hiver, disons entre le 1er janvier et le 1er avril [3 mois]. La probabilité, que l'on considérera constante, d'attraper un rhume chaque mois est de 48% (il s'agit bien évidemment d'un risque). Quel est le risque d'attraper un rhume durant l'unité de temps qu'est cette saison froide de 3 mois?  
 $\frac{0.48}{1/3} = 1.44$ . On peut donc s'attendre à attraper 1.44 rhume durant la période d'hiver. Il s'agit bien là d'un risque.
2. On passe une année en **vacances** dans une région, franchement pas très accueillante, où la probabilité de décéder chaque mois est évaluée à 33%. Quel est le risque de décéder pendant cette année?  
 $\frac{0.33}{1/12} = 3.96$

On le voit, cette mesure du risque peut donc être supérieure à 1. En soit cela ne pose pas de problème comme il s'agit d'un nombre moyen d'évènements espérés, mais pour des évènements qui ne peuvent pas se répéter, évènements dits *absorbants* (par définition la mortalité), l'interprétation n'est pas très intuitive.

Le risque étant constant dans chaque intervalle (mois), on peut prendre son inverse qui va mesurer la durée moyenne (espérée) jusqu'à l'occurrence de l'évènement.

<sup>5</sup>La relation  $h(t) = \frac{f(t)}{S(t)}$  et donc  $f(t) = h(t) \times S(t)$  est intéressante et importante car elle permet d'écrire la vraisemblance du processus probabiliste permettant d'estimer les paramètres des différentes analyses. On voit déjà sa proximité avec la fonction de masse de Bernoulli:  $f(y_i) = p^{y_i} \times (1 - p)^{1-y_i}$ . Se reporter à la section qui décrit la vraisemblance partielle de Cox pour s'en faire une idée plus précise.

Par analogie seulement, 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, ce risque inconditionnellement à la durée étant par définition égal à 1, mais jusqu'à quand on peut espérer (sur)vivre.

- Pour le rhume, la durée moyenne est de  $1.44^{-1} = 0.69$  du trimestre hivernal. On peut donc s'attendre, en moyenne, à attraper un rhume approximativement au début du mois de mars. Bien évidemment, certain.e.s attraperont un rhume avant, certaine.s après, certain.e.s aucun<sup>6</sup>.
- Pour l'année sabbatique, la durée moyenne de survie est de  $3.96^{-1} = 0.25$  d'une année soit 3 mois après l'arrivée dans la région.

### 4.5.1 Le risque cumulé $H(t)$

Le risque cumulé est égal à :

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

On peut alors réécrire toutes les autres quantités à partir de celle-ci:

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

*Exemple avec la loi exponentielle (risque constant)*

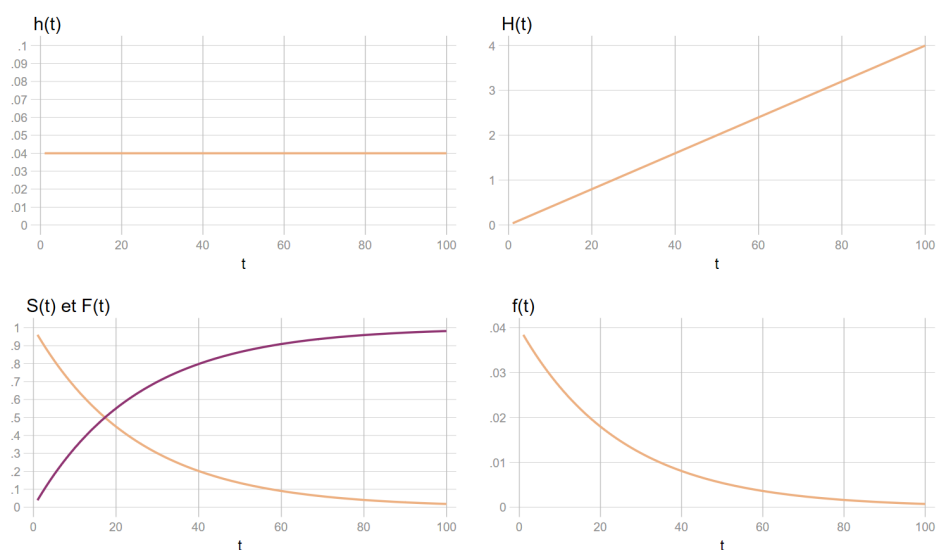
Si on pose que le risque (ou taux de risque) est strictement constant au cours du temps:  $h(t) = a$ , on dira qu'il suit une **loi exponentielle**. Cette situation est, par exemple, typique des processus dits sans mémoire comme la durée de vie des ampoules. Sans trop de difficulté, toutes les expressions peuvent être formellement définies:

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

---

<sup>6</sup>et une analyse plus fine pourrait s'intéresser sur l'effet du port du masque sur ce risque d'attraper le rhume

Figure 4.3: Grandeurs de la loi exponentielle avec  $h(t)=0.04$



### Exercice

- On a une population de 100 cochons d'Inde.
- On analyse leur mortalité (naturelle).
- Ici l'analyse est en temps discret.
- La durée représente le nombre d'année de vie.
- Il n'y a pas de censure ou troncature 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? Et quelle est sa valeur?

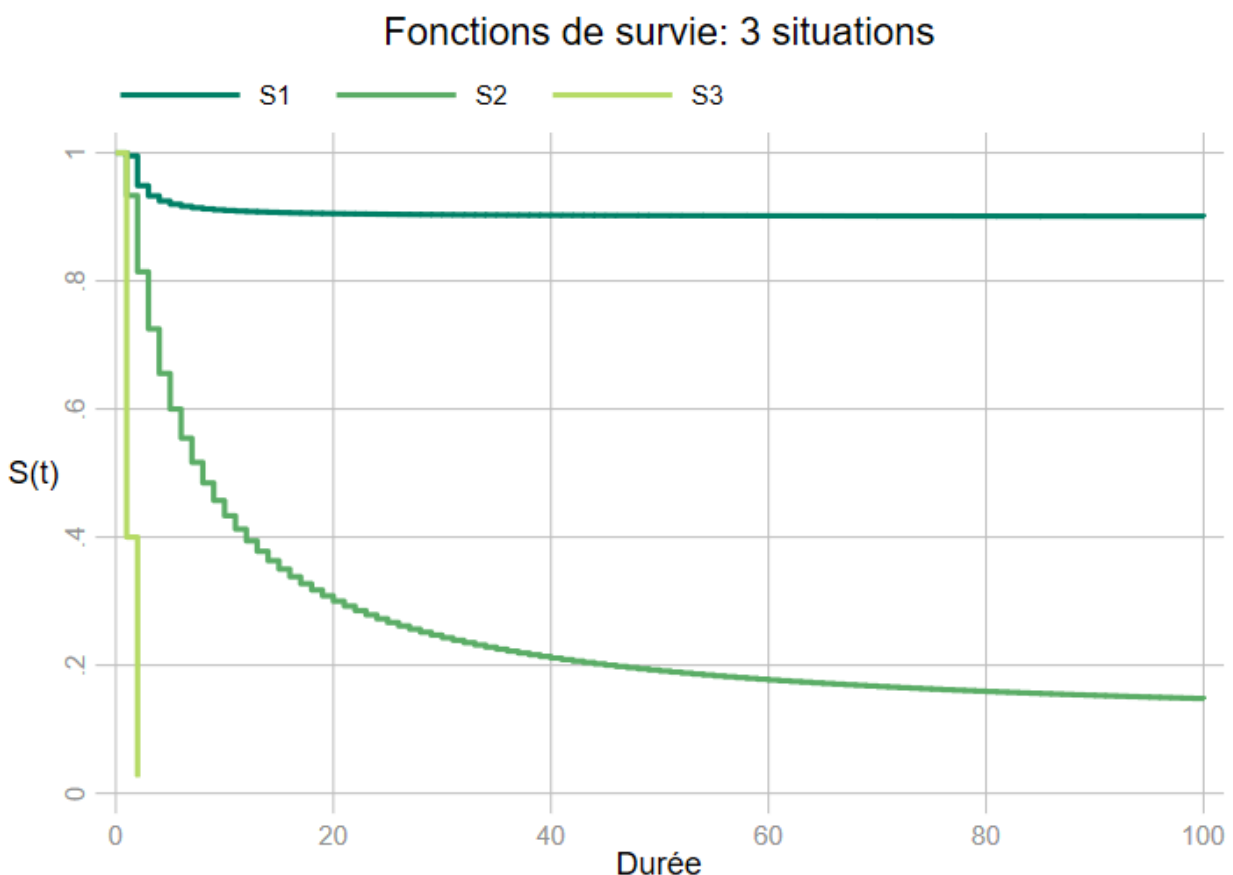
## 4.6 Remarques complémentaires

### 4.6.1 Formes typiques de la fonction de survie

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

Par anticipation, on peut déjà signaler que les fonctions de séjours qui sont représentées ci-dessous, font l'objet d'une estimation de type Kaplan-Meier.

Figure 4.4: Fonction de survie: 3 situation typiques... ou pas

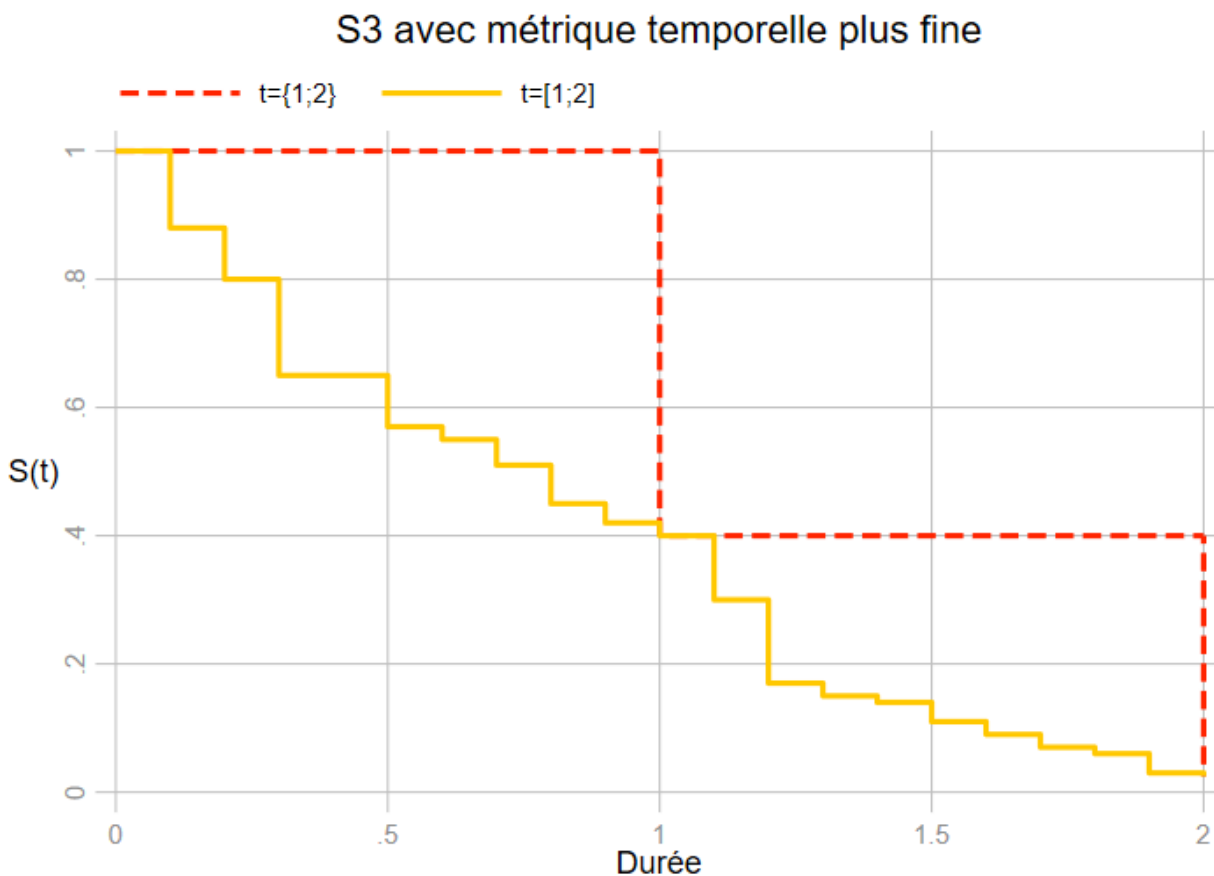


- **S1**: très peu d'évènements et la fonction de séjour suit une asymptote nettement supérieur à 0 ( $\lim_{t \rightarrow \infty} S(t) = a$  avec  $a > 0$ ). La question est plus délicate car on interroge l'exposition au risque d'une partie de l'échantillon ou, dit autrement on peut penser qu'une fraction est immunisée au risque. Cette problématique est rapidement posée en fin de formation.
- **S2**: la situation attendue
- **S3**: La survie tombe à 0 très/trop rapidement: il n'y a donc pas ou presque pas de durée (par exemple presque tout l'échantillon observe l'évènement la première année de l'exposition). 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 précédent  $S(t = 1) = 0.4$  ,  $S(t = 2) = 0.025$ , mais si on dispose par exemple de 10 points d’observations supplémentaires dans chaque intervalle groupé:

Figure 4.5: Fonction de survie et modification de la métrique temporelle



#### 4.6.2 Absence de censures à droites

Les méthodes qui vont être présentées plus tard **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. On peut alors utiliser d’autres méthodes plus, en première intentions plus simples, en analysant directement la distribution des évènements dans le temps. La durée ou une fonction de celle-ci pouvant être directement passées comme variable dépendante.

### 4.6.3 Utilisation des pondérations dans un schéma rétrospectif avec des biographies longues

#### ! Important

- Cette problématique mériterait au moins une section propre. Mais le sujet est compliqué et sensible. En attendant mieux, je rajoute quelques éléments.
- L'argumentaire écrit à la suite cet encadré est antérieur à septembre 2024. depuis des questionnements m'ont été adressés par l'équipe de groupe d'exploitation de l'enquête [Envie...](#) et il peuvent remettre un peu en cause cet argumentaire, la plage d'âges de l'enquête étant assez réduite (de 18 à 29 ans). Plus précisément, la question portait sur l'utilisation des pondérations dans l'estimation non paramétrique des courbes de survies.
- Pour ce qui est des modèles, en sciences sociales les variables entrant dans le calcul des pondérations sont généralement introduites dans les modèles, et à moins de revendre une analyse de type **Population average**, l'ajout des pondérations ne s'impose pas. Quel est le sens d'un *Population average* en analyse des durées lorsqu'on s'éloigne d'une plage d'observation très courte, comme un suivi sur 12 mois de reprises d'emploi de personnes au chômage ????
- Dans les domaines où s'appliquent régulièrement l'analyse des durées, à savoir la médecine et l'épidémiologie, la question des pondérations se pose lors d'études rétrospectives avec de fort soupçons d'effet de sélection (unbalanced data). En médecine on procède généralement à un recueil permettant d'appliquer directement des méthodes *case-control*, qui permet justement de contourner le problème. La problématique de la pondération ne se pose donc pas en terme de représentativité dans une population générale, mais en terme d'effet confondant, soit l'absence ou la mauvaise mesure hors redressement par la pondération d'une caractéristique corrélée à une covariable mais également à l'évènement qu'on analyse.
- Caractéristique de l'analyse des durées: la taille de l'échantillon se réduit au cours du temps par le jeu des censures et des évènements. Question: de quoi elles sont représentatives tout du long de la plage d'observation, alors quelles ont été calculées au moment de l'enquête, schéma classique pour les enquêtes en population générale.
- Dans la courte bibliographie accessible que j'ai donné en début de support, la question n'est pas traitée ni par G.Colletaz ni par S.Quantin (pourtant Insee)... Je marche seul!!!!
- [Note bien compliquée comme il le faut de T.Therneau et al... champ médecine](#)
  - Les informations permettant de calculer la pondération sont issus du recensement américain de 2000...le début du suivi commence au milieu des années 90. Mais l'analyse y est rétrospective. Les différentes techniques de pondération qui sont discutées cherchent à recalibrer l'échantillon pour contrôler le plus correctement possible la présence d'un effet de confusion avec l'âge. Il ne s'agit pas d'une analyse en population générale, ce qui n'aurait pas de sens vu qu'on analyse la survie à une maladie du sang.

[Eléments rédigés avant septembre 2024]

Une question assez récurrente concerne l'utilisation des poids de sondage dans les analyses de durées avec longueurs biographiques souvent assez longues. Leur utilisation ne me semble pas recommandée voire à exclure sauf exceptions. En effet les pondérations sont générées au moment de l'enquête, alors que les évènements étudiés peuvent remonter dans un passé plus ou moins lointain pour une partie de la population analysée. Si on regarde de plus près, la création de poids longitudinaux ne résoudrait pas grand chose, les pondérations devant être recalculées à chaque moment d'observation ou à chaque moment où des évènements se produisent. Par ailleurs on mélangerait régulièrement à un instant donné des personnes issues de générations tellement éloignées que cela rend impossible tout calage sur des caractéristiques d'une population à un instant  $t$ . 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?????

## **partie III**

# **Méthodes non paramétrique**

## 5 Estimations des fonctions de survie

Les méthodes non paramétriques portent généralement sur l'analyse des fonctions de survie ( $S(t)$ ) moins sur les fonctions de répartitions ( $F(t)$ )<sup>1</sup>, plus rarement sur les mesures d'incidence données par le risque cumulé. Deux méthodes d'estimations sont proposées : la méthode dite **actuarielle** et la méthode dite de **Kaplan & Meier**. Ces deux approches sont adaptées à des mesures différentes de la durée : plutôt discrète/groupée pour la technique actuarielle et plutôt continue pour Kaplan-Meier (KM). Cela induit un traitement différent de la censure dans l'estimation. La seconde est de très très loin la plus utilisée, médecine oblige, et en partie en raison des tests de comparaison, plus ou moins pertinents, qu'elle permet de réaliser.

### ! Important

- J'insiste sur la nécessité de passer par cette étape avant de se lancer *corps perdu* dans des modèles, comme ceux à durée discrète/groupée.
- Egaleme nt très utile, la comparaison graphique de courbes de séjour permet de repérer rapidement des violations fortes de l'hypothèse de proportionalité des risques, ou des situations de quasi *immunité*.
- Concernant les tests non paramétriques, ceux utilisant la technique du *logrank*, présentent beaucoup de défauts. Malheureusement encore très peu diffusée dans les sciences sociales, la comparaison des RMST (*Restricted Mean of Survival Time*), dérivée de l'estimateur de Kaplan Meier, me semble une solution largement supérieure, tant au niveau statistique qu'au niveau interprétatif.

### 5.1 Les fonctions de survie/séjour

#### 5.1.1 Les variables d'analyse

On a un échantillon aléatoire de  $n$  individus avec:

- Des indicateurs de fin d'épisode  $e_1, e_2, \dots, e_k$  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.
- En théorie, il ne peut pas y avoir d'évènement en  $t = 0$ .

---

<sup>1</sup>malheureusement, mais cela dépendra des options des logiciels

## 5.1.2 Calcul de la fonction de survie

Rappel: La fonction de survie donne la probabilité que l'évènement survienne après  $t_i$ , soit  $S(t_i) = P(T > t_i)$ . Pour survivre en  $t_i$ , il faut donc avoir survécu en  $t_{i-1}, t_{i-2}, \dots, t_1$ .

La fonction de survie renvoie donc des probabilités conditionnelles: on survit en  $t_i$  conditionnellement au fait d'y avoir survécu avant. Il s'agit donc d'un produit de probabilités.

Soit  $d_i = \sum e_i$  le nombre d'évènements observés en  $t_i$ , et  $r_i$  la population encore soumise au risque en  $i$ . On peut mesurer l'intensité de l'évènement en  $t_i$  en calculant le quotient  $q(t_i) = \frac{d_i}{r_i}$ .

Si le temps est strictement continu on devrait toujours avoir  $q(t_i) = \frac{1}{r_i}$ .

$S(t_i) = (1 - \frac{d_i}{r_i}) \times S(t_{i-1}) = S(t_i) = (1 - q(t_i)) \times S(t_{i-1})$ . En remplaçant  $S(t_{i-1})$  par sa valeur:  $S(t_i) = (1 - \frac{d_i}{r_i}) \times (1 - \frac{d_{i-1}}{r_{i-1}}) \times S(t_{i-2})$ .

Au final, en remplaçant toutes les expressions de la survie jusqu'en  $t_0$  ( $S(0) = 1$ ):

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

### i Application pour la suite du support

- On va analyser le risque de décéder (la survie) de personnes souffrant d'une insuffisance cardiaque. Le début de l'exposition est leur inscription dans un registre d'attente pour une greffe du coeur.
- Les covariables sont dans un premier temps toutes fixes: l'année (*year*) et l'âge (*age*) à l'entrée dans le registre, et le fait d'avoir été opéré pour un pontage aorto-coronarien avant l'inscription (*surgery*).
- Le début de l'exposition au risque est l'entrée dans le registre, la durée est mesurée en jour (*stime*). La variable évènement/censure est le décès (*died*). Les durées de la variable *stime* ont été regroupées par période de 30 jours pour réaliser des analyses à durée discrete. Cette nouvelle variable de durée a été appelé *mois*.
- L'introduction d'une dimension dynamique, la greffe, est donnée par les informations contenues dans les variables *transplant* et *wait*.
- La variable *compet* est une information simulée pour réaliser des analyses en risques concurrents.
- Les bases en format .csv, .sas7bdat<sup>2</sup> et .dta sont disponibles dans ce dépôt [\[lien\]](#)

Extrait de la base:

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

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

## 5.2 La méthode actuarielle

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

### 5.2.1 Estimation

#### 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:* en raison de la présence de censures à droite, le dernier intervalle étant ouvert jusqu'à la dernière sortie d'observation, il n'est pas conseillé de calculer des durées moyennes. On préfère utiliser la médiane ou tout autre quantile lorsqu'ils sont calculables.

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

---

<sup>2</sup>juste pour le souvenir

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

## R-Stata-Python-Sas

### 5.2.1.1 R

Les fonctions de survie avec la méthode dite actuarielle sont estimables avec le package **discSurv**. Avec le temps, il s'est étoffé, on peut maintenant paramétrer des intervalles (programmation pénible), mais les quantiles de la durée ne sont toujours pas estimables, ce qui est bien dommage, voire rend son utilisation peu intéressante.

### 5.2.1.2 Stata

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

### 5.2.1.3 Python

A l'heure actuelle, aucune fonction à ma connaissance.

### 5.2.1.4 Sas †

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

## 5.2.2 Application

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

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

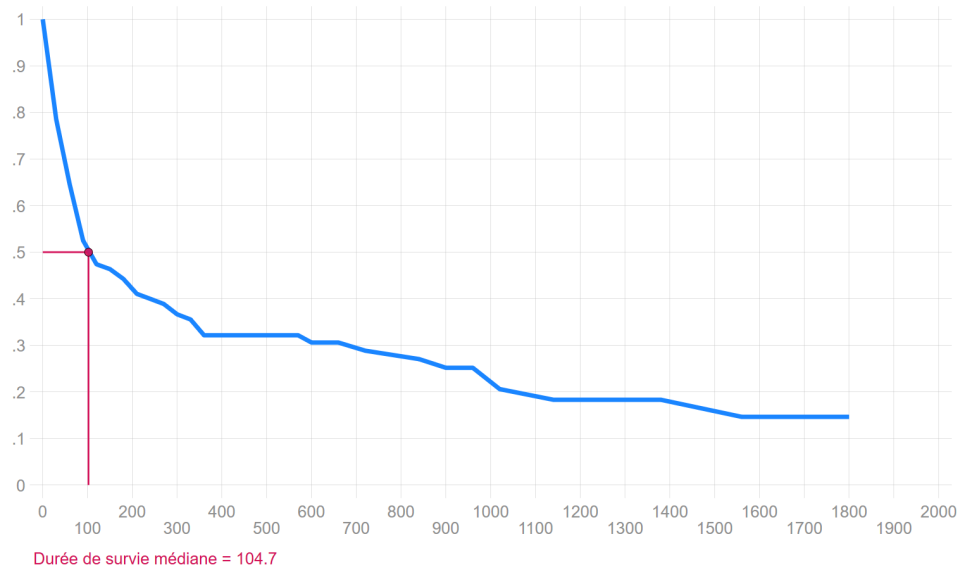


7.	180	210	.4425605	.3435417	.5368989
8.	210	240	.4105681	.3132064	.5052779
9.	240	270	.3997637	.3030412	.4945301
10.	270	300	.3888113	.2927645	.4836136
-----					
11.	300	330	.3665935	.2720434	.4613676
12.	330	360	.3554846	.2617823	.4501585
13.	360	390	.3216289	.2308275	.4157428
14.	390	420	.3216289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

Table 5.2: Quantiles de la fonction de séjour type actuarielle - Bornes Sas

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

Figure 5.1: Courbe de survie: estimation méthode actuarielle



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

## 5.3 La méthode de Kaplan-Meier

- L'approche qui exploite toute l'information disponible est celle dite de **Kaplan-Meier** (*KM*).
- Il y a autant d'intervalles que de durées où l'on observe au moins un évènement.
- Au lieu d'utiliser des intervalles prédéterminés, l'estimateur KM va définir un intervalle entre chaque évènement enregistré.
- La fonction de survie estimée par la méthode KM est une fonction en escalier (stairstep), d'où une estimation dite "discrete".
- 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.

### 5.3.1 Estimation

#### Définition du Risk Set ( $r_i$ )

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

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

## Calcul de $q_i$

On applique la méthode de la section précédente avec:

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

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

## Récupération de la médiane

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

La définition retenue est donc conventionnelle. On va prendre la valeur de la durée qui se situe juste **en dessous** de 50% de survivant.e.s. Elle est donc définie tel que  $S(t) \leq 0.5$ . Attention, il n'est pas impossible que le % de survivant.e.s soit bien en deçà de 50% pour l'obtention cette durée médiane.

## R-Stata-Python-Sas

### 5.3.1.1 R

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

### 5.3.1.2 Stata

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

### 5.3.1.3 Python

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

### 5.3.1.4 SAS†

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

## 5.3.2 Application

On reprend l'exemple précédent.

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

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

[Résultats non reportés à partir de t=1000 ]

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

Table 5.3: Quantiles de la fonction de séjour type Kaplan-Meier

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

Figure 5.2: Courbe de survie: estimation méthode actuarielle

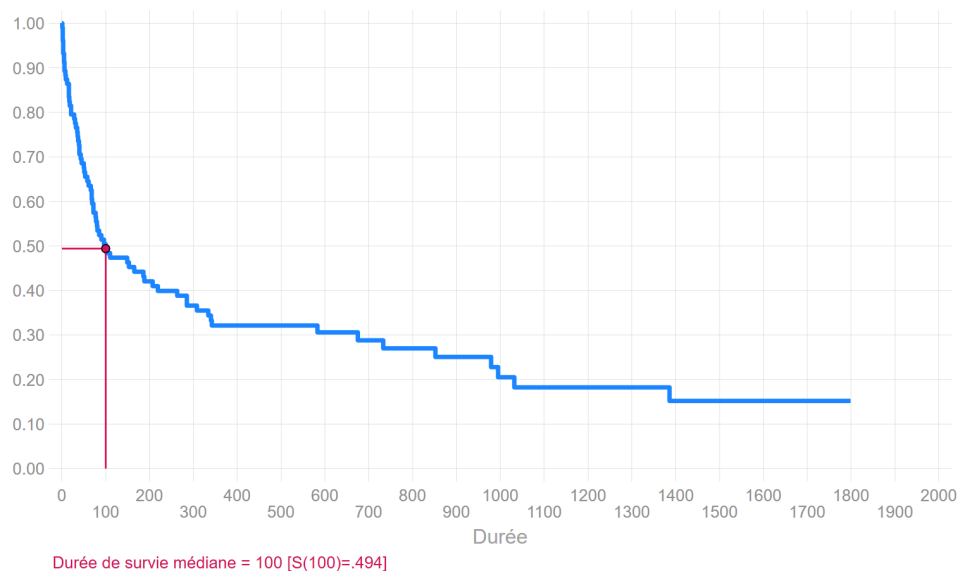
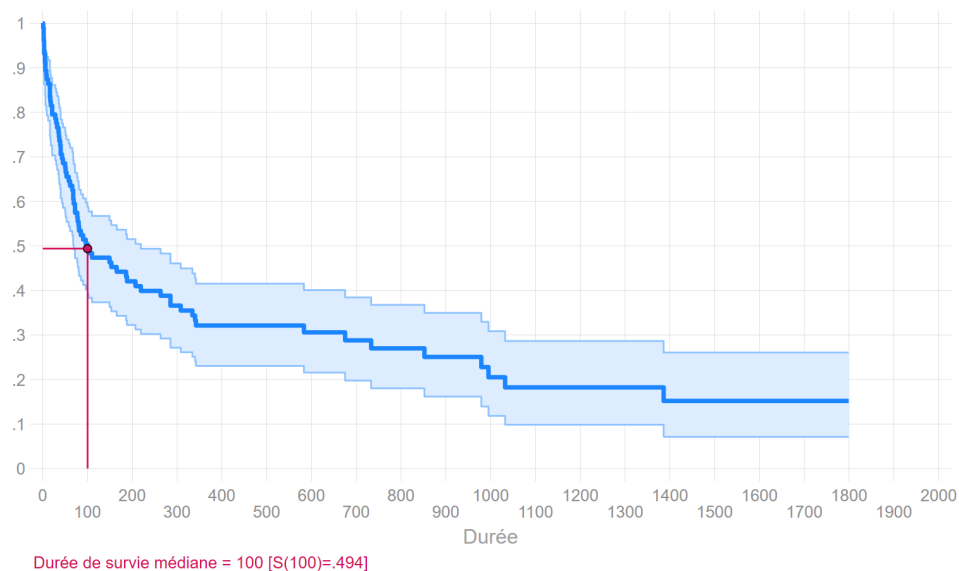


Figure 5.3: Courbe de survie: estimation méthode actuarielle + CI



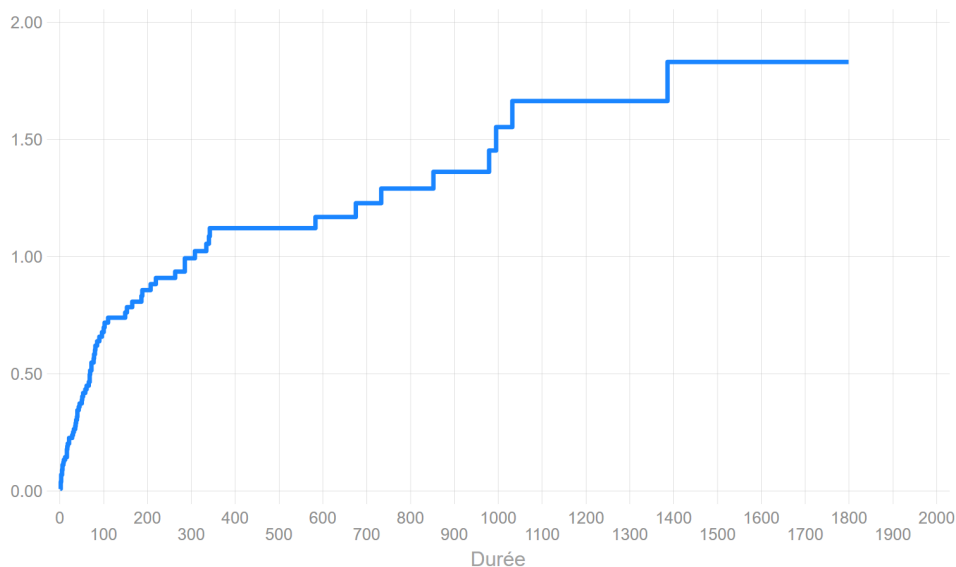
### 5.3.3 Quantités associées à l'estimateur Kaplan-Meier..

**Le risque cumulé:** estimateur de Nelson Aalen

Il est simplement égal à:

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

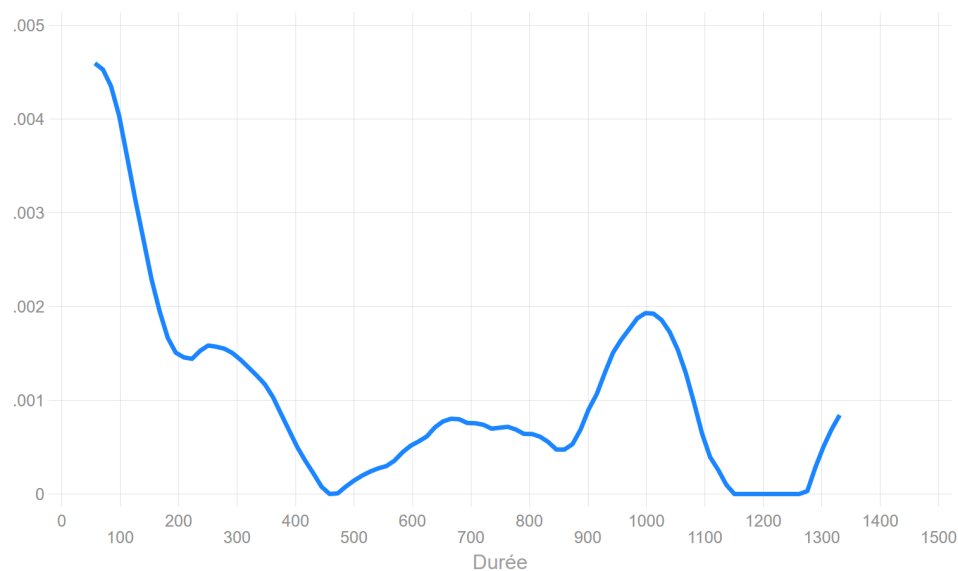
Figure 5.4: Risque cumulé: estimateur Nelson-Aalen



#### Le risque ou taux de hasard instantané

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

Figure 5.5: Risque instantané: estimateur du Kernel



**i** Note

Il n'est pas inutile de noter qu'il n'y a pas de *formule* toute faite pour obtenir des valeurs du risque instantané. Ce type de méthode par lissage est pleinement paramétrable, par exemple sa fenêtre, ce qui implique que son profil varie d'un paramétrage à l'autre. Le graphique précédent a été fait avec Stata, si on utilisait le package `mu` on visualiserait une courbe différente... mais on pourrait retrouver celle qui est représenté ici en changeant les paramètres de lissage.



## 6 Tests de comparaison

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

Deux méthodes sont utilisées:

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

### 6.1 Tests du log-rank

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

Dans leur logique, ces tests entrent dans le cadre des tests d'indépendance du Khi2, même si formellement ils relèvent des techniques dites de **rang** (d'où le nom).

Il s'agira donc de comparer des effectifs observés à des effectifs espérés à chaque moment d'évènement. La principale différence réside dans le calcul de la variance de la statistique du test qui, ici, suit assez logiquement une loi hypergéométrique [proche loi binomiale mais avec tirage avec remise].

#### 6.1.1 Principe de calcul de la statistique de test

- **Effectifs observés en  $t_i$ :**  $o_{i1}$  et  $o_{i2}$  sont égaux à  $d_{i1}$  et  $d_{i2}$ , et leur somme pour tous les temps d'évènement à  $O_1$  et  $O_2$ .
- **Effectifs espérés** (hypothèse nulle  $H_0$ ): comme pour une statistique du  $\chi^2$  on se base sur les marges, avec le risque set ( $R_i$ ) en  $t_i$  pour dénombrer les effectifs, soit  $e_{i1} = R_{i1} \times \frac{d_i}{R_i}$  et  $e_{i2} = R_{i2} \times \frac{d_i}{R_i}$ . Leur somme pour tous les temps d'évènement est égale à  $E_1$  et  $E_2$ . Le principe de calcul des effectifs observés reposent donc sur l'hypothèse d'un rapport des risques toujours égal à 1 au cours du temps (*hypothèse fondamentale de risques proportionnels*).
- **Statistique du log-rank:**  $(O_1 - E_1) = -(O_2 - E_2)$ .

- **Statistique de test:** sous  $H_0$ ,  $\frac{(O_1 - E_1)^2}{\sum v_i}$ , avec  $v_i$  la variance de  $(o_{i1} - e_{i2})$ , suis un  $\chi^2(1)$ . Si on teste simultanément la différence de  $g$  fonctions de survie, ce qui n'est pas une bonne idée en passant, la statistique de test suis un  $\chi^2(g - 1)$ .

## 6.1.2 Les principaux tests log-rank

Le principe de construction des effectifs observés et espérés reste strictement 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 encore soumise au risque à chaque durée où au moins un évènement est observé.

- **Test du log-rank:**  $w_i = 1$   
Il accorde le même poids à toutes les durées d'évènement. C'est le test standard, le plus utilisé. Il n'y a donc pas de pondération au final.
- **Test de Wilconxon-Breslow-Grehan:**  $w_i = R_i$   
Les écarts entre effectifs observés et espérés sont pondérés par la population soumise à risque en  $t_i$ . Le test accorde plus de poids au début de la période analysée, et il est sensible aux différences de distributions entre les strates des observations censurées.
- **Test de Tarone-Ware:**  $w_i = \sqrt{R_i}$   
Variante du test précédent, il atténue le poids accordé aux évènements au début de la période d'observation. Il est par ailleurs moins sensible au problème de la distribution des censures entre les strates.
- **Test de Peto-Peto :**  $w_i = S_i$   
La pondération est une variante de la fonction de survie KM (avec  $R_i = R_i + 1$ ). Le test n'est pas sensible au problème de distribution des censures.
- **Test de Fleming-Harington:**  $w_i = (S_i)^p \times (1 - S_i)^q$  avec  $0 \leq p \leq 1$  Il permet de paramétrer le poids accordé au début où à la fin de temps d'observation. Si  $p = q = 0$  on retrouve le test de base non pondéré<sup>1</sup>.

### En pratique/remarques:

- Les tests du log-rank sont sensibles à l'hypothèse de risques proportionnels (voir **modèle semi-paramétrique de Cox**). En pratique si des courbes de séjours se croisent, il est fortement déconseillé de les utiliser analytiquement. Cela ne signifie pas que si les courbes ne se croisent pas, l'hypothèse de proportionnalité des risques est respectée : des rapports de risque peuvent au cours du temps s'intensifier, se réduire ou, le cas échéant s'inverser, ce qui est typique d'un croisement.
- Effectuer un test global (multiple/omnibus) sur un nombre important de groupes (ou  $>2$ ) peut rendre les p-value artificiellement très basses (idem test classique d'indépendance). Il peut être intéressant de tester des courbes deux à deux (idem qu'une régression avec covariable discrète), en conservant un seul degré de liberté. Des méthodes de correction du test multiple sont possibles ou disponibles si on utilise R, bien que la bonne méthode n'a jamais fait consensus chez les statisticiens.

---

<sup>1</sup>A ma connaissance il n'est pas implémenté dans R

### Note

Cette question de test multiple ne se pose pas dans le support. Mais c'est le cas avec la formation *en live* pour les données traitées dans les TP

## R-Stata-Python-Sas

### 6.1.2.1 R

On utilise la fonction **survdif** de la librairie **survival**. Le résultat du test de Peto-Peto est affiché par défaut (**rho=1**). Si on souhaite utiliser le test non pondéré, on ajoute l'option **rho=0**. Pour obtenir le résultat d'un test multiple corrigé (plus d'un degré de liberté), on peut utiliser la fonction **pairwise\_survdif** du package **survminer**. Cette fonction permet également d'obtenir des tests 2 à 2.

Je conseille de rester sur l'option **Peto-Peto** et dans le cas d'une variable à plus de deux modalités, d'utiliser la fonction de **survminer** **pairwise\_survdif**.

### 6.1.2.2 Stata

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

### 6.1.2.3 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.survdif** de **statmodels** (non pondéré, Wilcoxon - appelé ici Breslow- et Tarone-Ware).

### 6.1.2.4 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**.

## 6.1.3 Application

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

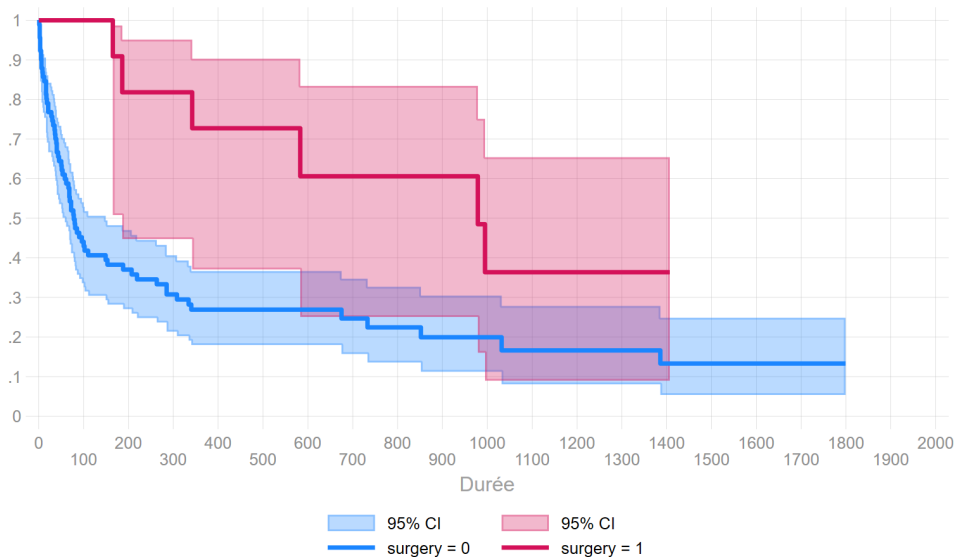


Table 6.1: Résultats des tests du logrank

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

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

## 6.2 Comparaison des RMST

RMST: *Restricted Mean of Survival Time*

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

### Principe

- L'aire sous la fonction de survie représente la durée moyenne d'attente jusqu'à l'évènement, soit une espérance de survie.
- En présence de censure à droite, il faut borner la durée maximale  $t^* < \infty$ . Cette espérance de survie ou de séjour s'interprète donc sur un horizon fini. On est très proche d'une mesure en analyse démographique type « espérance de vie partielle ».

- $RMST = \int_0^{t^*} S(t)dt$ .
- On peut facilement comparer les RMST de deux groupes, en termes de différence ou de ratio.
- Par défaut on définit généralement  $t^*$  à partir le temps du dernier évènement observé. Il est néanmoins possible de calculer le RMST sur des intervalles plus court, ce qui lui permet une véritable souplesse au niveau de l'analyse.

## R-Stata-Python-Sas

Attention, selon les logiciels la durée max par défaut n'est pas la même. Pour R et Sas, il s'agit du dernier évènement observé sur l'ensemble de l'échantillon, alors que Stata prend la durée qui correspond au dernier évènement observé le plus court des deux groupes . Cela affectera légèrement la valeur des Rmst estimées par défaut.

Pour l'exemple, la durée maximale utilisée par R est de 1407 jours alors que pour Stata elle est de 995 jours.

### 6.2.0.1 R

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

### 6.2.0.2 Stata

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

### 6.2.0.3 Python

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

### 6.2.0.4 SAS†

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

## Application

Avec  $tmax = 1407$ :

Table 6.2: Estimation des Rmst pour la variable surgery

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

Table 6.3: Différences entre Rmst pour la variable surgery

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

Ici  $t^*$  est égal à 1407 jours, soit la durée qui correspond au dernier décès observé.

Sur un horizon de 1407 jours, ces individus opérés d'un pontage peuvent espérer vivre 884 jours en moyenne, contre 379 jours pour les autres. La durée moyenne de survie est donc 2.3 fois plus importante pour les personnes opérées (rapport des Rmst = 2.3 ), ce qui correspond à une différence de 379 jours.

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

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

#### Note

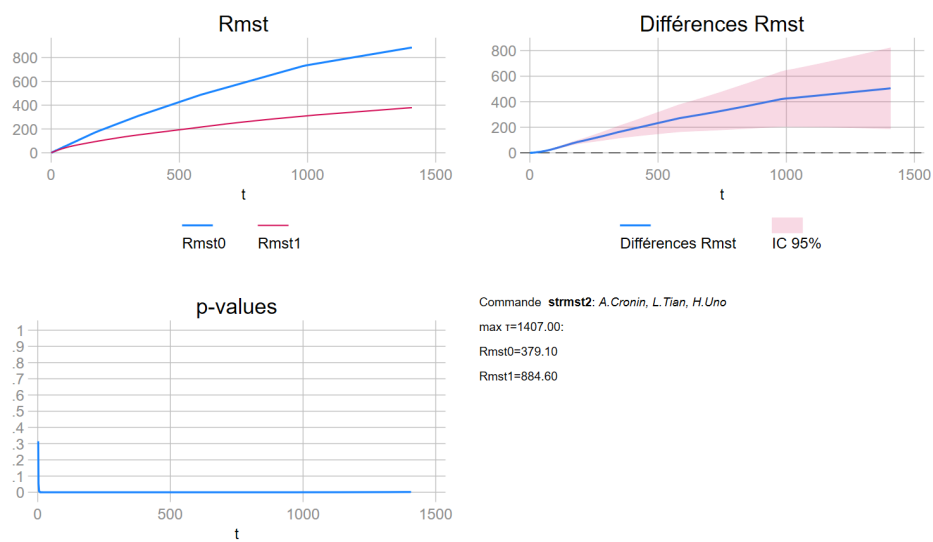
Pour la version pdf, seulement une grosse dizaine de points a été sélectionné en raison de la longueur du tableau.

# Seulement une partie de l'output: se reporter à la version html

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

	995	734.7576	310.1678	424.5898	204.0643	645.1152	.0001609	
	1032	748.2121	317.5443	430.6678	202.7468	658.5889	.0002127	
	-----							
	1141	787.8485	335.6531	452.1953	200.7097	703.681	.0004248	
	1321	853.303	365.5577	487.7454	191.5434	783.9473	.0012492	
	1386	876.9394	376.3565	500.5829	186.9499	814.2158	.0017585	
	1400	882.0303	378.2173	503.813	186.4392	821.1869	.0018625	
	1407	884.5757	379.1476	505.4281	186.1745	824.6817	.0019162	
	-----							
	-----							

Figure 6.1: Comparaison des Rmst à chaque jour où au moins un décès est observé



## **partie IV**

# **Modèles à risques proportionnels**



# 7 Introduction aux modèles

## 7.1 Proportionnalité des risques

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

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

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

### Le risque de base

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

### Risques proportionnels

Cette hypothèse stipule l'invariance dans la durée du *rapport des risques* (**hazard ratio**).

Exemple:

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

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

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

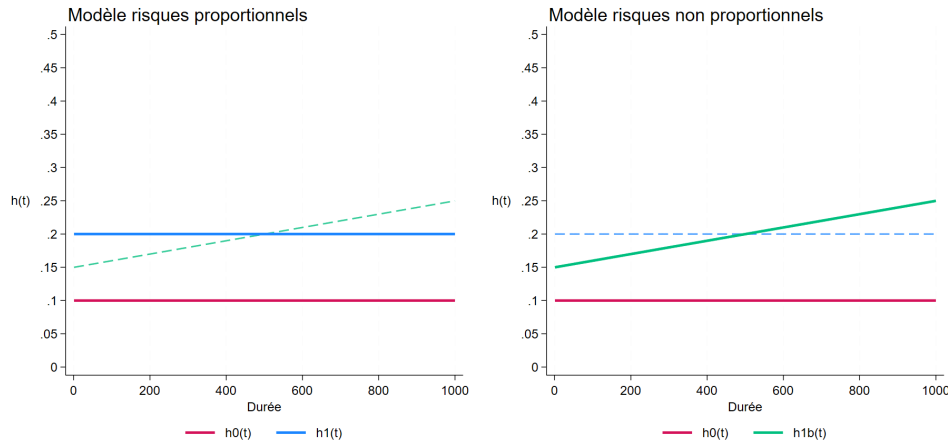
$$\frac{h_A(t)}{h_B(t)} = \frac{e^{bX_A}}{e^{bX_B}} = e^{b(X_A - X_B)}$$

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

---

<sup>1</sup>On rappellera qu'en durée continue, seule positivité du risque doit être assurée, d'où l'expression *hazard rate*

Figure 7.1: L'hypothèse de proportionalité des risques



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

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

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

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

## 7.2 Les modèles

- **Modèle semi-paramétrique de Cox (1972)**

Le modèle estime directement les  $b$  indépendamment de  $h_0(t)$ . C'est pour cela qu'il est appelé modèle **semi-paramétrique de Cox**. Les rapports de risque ( $e^b$ ) seront utilisés dans un deuxième temps pour estimer la baseline  $h_0(t)$ , qui peut s'avérer nécessaire pour calculer des fonctions de survie ajustées. Le respect de l'hypothèse de proportionnalité est donc importante et doit donc être analysée.

- **Modèle à durée discrète/groupée** Sa spécification diffère quelque peu de la présentation usuelle d'un modèle à risque proportionnel. Toutefois, il est régi par une hypothèse de proportionnalité. Le non respect de l'hypothèse est moins critique car la baseline du taux de risque est estimée simultanément aux autres paramètres. Il est comme son nom l'indique, particulièrement adapté aux durées discrètes ou groupées. Avec une spécification logistique, les Odds vont sous certaines conditions (souvent respectée), se confondre avec des probabilités/risques. Lorsque le nombre de points d'observations ( $t$ ) n'est pas trop faible, les résultats obtenus sont très proches de ceux issus

directement d'un modèle de Cox. On peut souligner qu'une première version de ce modèle a été à l'origine proposé par Cox lui même peu de temps après son modèle semi-paramétrique.

- **Les modèles paramétriques standards**

Les modèles dits de Weibull, exponentiel, Gompertz ont une spécification sous hypothèse de risque proportionnel. Ils seront traités brièvement dans les compléments (ça peut faire grincer des dents chez les épidémio). Historiquement, le modèle de Cox est une réponse à une possible difficulté dans l'ajustement du risque par une loi de distribution du risque a priori. Il est donc déjà important de noter que le modèle de Cox, devenu un gold standard<sup>2</sup> n'était à l'origine qu'une stratégie de repli.

- **Modèle paramétrique de Parmar-Royston**  $h_0(t)$ , via le risque cumulé  $H(t)$ , est estimé simultanément avec les rapports de risques en utilisant la méthode des *splines cubiques*. Il est maintenant implémenté dans les logiciels standards (R, Stata, Sas). Les rapports de risque obtenus sont très proches de ceux estimés par le modèle classique de Cox. Il offre donc une alternative surement intéressante au Cox standard, et il s'est maintenant largement diffusé dans l'analyse des effets cliniques.
- **Modèle à non proportionnalité**: on a bien évidemment les modèles paramétriques de type *AFT* (Accelerated Failure Time), le modèle à *pseudo observations* d'Andersen, sur le papier très séduisant mais qui butera régulièrement en sciences sociales sur l'hypothèse de non corrélation entre censures à droite et covariable. Dans le domaine du machine learning, à visée prédictive, il y a depuis son origine une version modèle de survie dans les *forêts aléatoires*. Malheureusement je n'ai jamais pu ou voulu m'y consacrer.

---

<sup>2</sup>Au regret de Cox lui même

# 8 Le modèle de Cox

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

## 8.1 Le modèle

### 8.1.1 La vraisemblance partielle et estimation des paramètres

On se situe dans une situation où la durée est mesurée sur une échelle strictement continue. Il ne peut donc y avoir qu'un seul événement observé en  $t_i$ .

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

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

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

#### Vraisemblance partielle de Cox

Comme  $f(t_i) = h(t_i) \times S(t_i)$ <sup>1</sup>, 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 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:  $\sum_{j \in R_i} h(t_i)$ , où  $j \in R_i$  est l'ensemble des observations soumises au risque en  $t_i$ .

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

La vraisemblance partielle retient seulement le premier terme de la vraisemblance<sup>2</sup>, soit:

<sup>1</sup>Se reporter à la définition des grandeurs dans la section *Théorie*

<sup>2</sup>Dans un entretien en 1994, Cox a précisé qu'il avait cette manipulation pour voir, sans idée préconçue

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

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

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

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

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

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

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

*Condition nécessaire: pas d'évènement simultané:* en présence d'évènements mesurés simultanément, l'estimation de la vraisemblance doit faire l'objet d'une correction.

*Correction de la vraisemblance avec des évènements simultanés:*

- La **méthode dite exacte**: Comme il ne doit pas y avoir d'évènement simultané, on va introduire à la vraisemblance partielle toutes les permutations possibles des évènements observés au même moment. Bien qu'en  $t_i$  on observe au même moment l'évènement pour 2 observations (A,B) une métrique temporelle plus précise permettrait de savoir si A s'est produit avant B ou B s'est produit avant A (2 permutations). Comme le nombre de permutations est calculé à l'aide d'une factorielle <sup>3</sup>, avec 3 évènements mesurés simultanément, on obtient 6 permutations ( $3 \times 2 \times 1$ ). Problème: le nombre de permutations pour chaque  $t_i$  peut devenir très vite particulièrement élevé. Par exemple pour 10 évènements simultanés, le nombre de permutations est égal à 3,628,800. Le temps de calcul devient extrêmement long, et ce type de correction totalement inopérant.
- La **méthode dite de Breslow**: il s'agit d'une approximation de la méthode exacte permettant de ne pas avoir à intégrer chaque permutation. *Cette approximation est utilisée par défaut par les logiciels Sas et Stata.*

---

<sup>3</sup> $n! = (n) \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$

- La **méthode dite d'Efron**: elle corrige l'approximation de Breslow, et est jugée plus proche de la méthode exacte. *C'est la méthode utilisée par défaut avec le logiciel R*, et elle est disponible avec les autres applications.

## 8.1.2 Estimation des paramètres

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

- Conditions de premier ordre: calcul des équations de score à partir des dérivées partielles. Solution:  $\frac{\partial \log(PL)}{\partial b_k} = 0$ . On ne peut pas obtenir de solution numérique directe.

Remarque: les équations de score seront utilisées pour tester la validité de l'hypothèse de constance des rapports de risque pour calculer les **résidus de Schoenfeld**... Le grand sujet (voir plus loin).

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

### Éléments de calcul

En logarithme (sans évènement simultané), la vraisemblance partielle s'écrit:

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

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

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

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

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

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

Cette expression va permettre d'analyser le respect ou non de l'hypothèse de risques proportionnels via les *résidus de Schoenfeld*.

### 8.1.3 Lecture des résultats

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

**Covariable binaire (indicatrice)**  $X = (0, 1)$ :  $RR = \frac{h(t | X=1)}{h(t | X=0)} = e^b$ .

A chaque moment de la durée  $t$  (jour, mois, années...), le risque d'observer l'évènement est  $e^b$  fois plus important/plus faible pour  $X = 1$  que pour  $X = 0$ .

**Covariable mesurée (quantitativement)** (mais fixe dans le temps)

$RR = \frac{h(t | X=a+c)}{h(t | X=a)} = e^{c \times b}$ . On prendra pour illustrer une variable type âge au début de l'exposition au risque ( $a$ ) et un delta de comparaison avec un âge inférieur  $c$ .

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

#### Exemple pour les insuffisances cardiaques

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

Table 8.1: Cox: log Hazard Ratio (Risks Ratio)

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

Table 8.2: Cox: Hazard Ratio (Risks Ratio)

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

On retrouve la même lecture que lors des tests non paramétriques pour l'opération, à savoir qu'un pontage réduit les risques journaliers de décès pendant la période d'observation (ie augmente la durée de survie).

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

Les résultats apparaissent, sans connaissance médicale très poussée, conforme à ce qu'on pourrait attendre.

### 8.1.4 R

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

#### 8.1.4.1 Stata

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

#### 8.1.4.2 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**.

#### 8.1.4.3 SAS†

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

## 8.2 Analyse de la constance des rapports de risque

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



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

Ce test a été proposé par P.Grambsch et T.Therneau <sup>4</sup> dans un cadre à durée strictement continue. Il repose originellement sur une régression linéaire estimée avec les moindres carrés généralisés (GLS) correction de l'autocorrélation des erreurs avec des séries). Dans un premier temps pour des raisons plutôt pratiques (informatique), le test a une version moindres carrés ordinaires (OLS). Jusqu'en 2020, tous les logiciels ne proposaient que le test OLS (Sas, Stata, Python et Survival v2). T.Therneau avec la Version 3 de package **survival** a substitué - assez brutalement - un test GLS<sup>5</sup> au test OLS. Et cela pose de gros problèmes, en particulier en sciences sociales. Ceci a été montré de manière très convaincante par [Shawna K. Metzger](#).

On doit également souligner que pour P.Grambsch et T.Therneau <sup>6</sup> n'est qu'un moyen parmi d'autres d'analyser une violation de l'hypothèse de proportionnalité. Ce n'est pas *the solution* (comme tout autre test au passage). Le croisement des courbes de séjours peut-être suffisant pour alerter sur cette violation. Ou tout simplement bien regarder les données qu'on a en face des yeux.... par exemple pour la variable *surgery*, il y a-t-il besoin de faire un test? Regardez bien.

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

### *Principe de calcul des résidus*

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

Pour une observation dont l'évènement s'est produit en  $t_i$ , le *résidu brut de Schoenfeld* pour la covariable  $X_k$ , après estimation du modèle, est égal à:

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

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

---

<sup>4</sup>Il s'agit bien de la personne qui maintient le package **survival** dans R

<sup>5</sup>Moindres carrés généralisés

<sup>6</sup>Se reporter à leur ouvrage *Modeling Survival Data: Extending the Cox Model* (2001)

Pourquoi?

Par l'exemple, sans censure à droite et en ne considérant que les résidus bruts:

Avec un rapport de risque strictement égal à 1 en début d'exposition, une population soumise au risque ( $R_0 = 100$ ) avec 50 hommes et 50 femmes. 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 l'espérance d'avoir la caractéristique  $X$  est toujours égal à 0.5 et les résidus bruts prendront toujours la valeur -.5 si  $X = 0$  et .5 si  $X = 1$ . En faisant une simple régression linéaire entre les résidus, qui alternent ces deux valeurs, et  $t$ , le coefficient estimé sera en toute logique très proche de 0.

De manière encore plus simple, cette proportionnalité avec un risque ratio égal à 1 suggère qu'au cours de la durée d'observation, on observe une succession d'un même nombre d'hommes et de femmes qui connaissent l'évènement.

En revanche, si tous les hommes ou presque avaient observés l'évènement plutôt en début d'exposition et si toutes les femmes ou presque avaient observé l'évènement plutôt en fin d'exposition, l'hypothèse de proportionnalité pourraient fortement remise en cause. Et on obtiendrait au final un rapport de risque moyen égal à 1.

On trouvera des éléments de calcul du test OLS [ici](#)

#### Avertissement

- **Test omnibus:** Ne pas l'utiliser bien qu'il figure généralement en bas des output. Il n'a pas d'interprétation directe, et les p-values peuvent présenter des valeurs très faibles alors que ce n'est pas le cas pour les covariables prises une à une. Rester si possible le cas à un test à un degré de liberté, variable par variable.
- **Transformations de la durée:** n'importe quelle fonction de la durée peut être utilisée pour réaliser le test. On retient généralement les fonctions suivantes:  $g(t) = t$  (« identity »),  $g(t) = \log(t)$ ,  $g(t) = KM(t)$  ou  $g(t) = 1 - S(t)$  où  $S(t)$  est l'estimateur de Kaplan-Meier. Enfin une transformation appelée « rank » est utilisée seulement pour les durées strictement continue ou suffisamment dispersées. Par exemple  $t = (0.1, 0.5, 1, 2.6, 3)$  donne une transformation  $t = (1, 2, 3, 4, 5)$ . A savoir :  $g(t)=t$  rend le test relativement sensible aux évènements tardifs lorsque la population restant soumise est peu nombreuse (outliers).
- Par défaut Stata, Sas, Python:  $g(t) = t$
- Par défaut R:  $g(t) = 1 - S(t)$

#### TRES IMPORTANT: Pourquoi la version exacte (GLS) du test pose problème

- Il s'agit de la version d'origine du test (1994) mais implémenté seulement en 2020 dans la version 3 du package `survival` de R. Les autres logiciels ne l'implémentent pas.
- S.Metzger a effectué une décomposition complète de sa statistique dans sa version exacte [lien pour les plus acharné.e.s](#) et l'a comparé à la version appelé "approximation" (je préfère "version Ols"). Résultats:

- Un terme *résiduel* se situant au numérateur de la statistique, qui traduit le degré de corrélation entre les covariables est présent. Ce qui signifie:
- Qu'en présence de corrélations entre covariables, même limitées, la statistique du test augmente artificiellement, et produira tout aussi artificiellement mais mécaniquement des p-values très faibles. Le test exact s'écarte alors fortement du test OLS, alors qu'il ne devrait être juste qu'un peu plus précis. Si les travaux de S.Metzger s'appuient bien sur des simulations, ses résultats sont vérifiés avec des données réelles qu'elle a également confrontée. C'est également ici le cas avec les données du support comme celles des TP. Pour la base utilisée dans ce support, c'est la corrélation entre les variables *surgery* et *year* qui explique les fortes différences pour la deuxième variable entre les deux versions du test.
- Cela signifie également que le test, qui consiste à synthétiser l'approche par l'interaction (voir plus bas), donnera des résultats contradictoires avec cette dernière.
- Dans R, et seulement dans R:
  - T.Therneau, propriétaire du package **survival** et visiblement discutant du travail de S.Metzger, n'a toujours pas remis le test OLS dans le package, ce qui est quand même problématique au niveau reproductibilité et réplicabilité dans le temps et dans l'espace des logiciels. C'est fâcheux.
  - Ne pas paniquer, j'ai récupéré le test OLS et on peut le charger et l'utiliser très facilement. Voir la section programmation dédiée à R pour le détail de la démarche (ou directement ce [lien](#)).

On ne présentera donc ici que la version OLS (idem Stata, Python, Sas **†**).

### Test OLS avec $g(t) = t$

Table 8.3: Test OLS Grambsch-Therneau avec  $g(t) = t$

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

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

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

Table 8.4: Test Grambsch-Therneau avec  $g(t) = 1 - S(t)$

Variables	chi2	df	P>Chi2
year	1.96	1	0.162

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

## R-Stata-Python-Sas

### 8.2.1.1 R

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

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

### 8.2.1.2 Stata

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

### 8.2.1.3 Python

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

### 8.2.1.4 SAS†

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

## 8.2.2 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 aux moment où au moins un évènement a été observé.

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

Table 8.5: Base spittées sur les intervals d'évènement

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

Les bornes des intervalles  $[t_0; t]$  présentent des valeurs seulement lorsqu'un évènement s'est produit (principe de la vraisemblance partielle). Il n'y a donc pas de valeurs pour  $t$  et  $t_0$  en  $t = 4$  pour  $id = (2, 3)$  et  $t = 7, 10, 11, 13, 14, 15$  pour  $id = 3$ .

Les deux individus observent l'évènement en  $t = 6$  pour  $id = 2$ , et en  $t = 16$  pour  $id = 3$ . Avant ce moment la valeur de la variable évènement/censure (ici  $d$ ) prend toujours la valeur 0, et prend la valeur 1 le jour du décès.

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

*Introduction d'une interaction avec une fonction de la durée*

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

Le modèle s'écrit:

$$h(t|X, t) = h_0(t)e^{b_1age+b_2year+b_3surgery+b_4(surgery \times t)}$$

Le modèle avec cette interaction donne les résultats suivants:

Table 8.6: Modèle de Cox avec une interaction entre une fonction de la durée et la variable \*surgery

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

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

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

### A noter

- Le modèle n'est plus un modèle à risque proportionnel. La variable *surgery* n'est plus une variable **fixe** mais une variable tronquée dynamique qui prend la valeur de  $t$  pour les personnes qui ont été opérées d'un pontage avant leur entrée dans le registre de greffe.

Si *surgery* = 0

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

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

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

Exemple pour une variable quantitative (*age*)

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

- L'altération des rapports de risque dépend de la forme fonctionnelle de l'interaction choisie. Ici la variation dans la durée du rapport des risque est constante, ce qui est une hypothèse assez forte. On a, en quelques sorte, réintroduit une hypothèse de proportionnalité, ici sur le degré d'altération des écarts de risques dans le temps, qui devient lui même strictement constant.

## 8.2.3 Que faire ?

### *Ne rien faire*

On interprète le risque ratio comme un ratio moyen pendant la durée d'observation (P.Allison), c'est le principe même du modèle. Difficilement soutenable pour l'analyse des effets cliniques, il faut reconnaître que ceci peut être envisagé dans d'autres domaines, comme en sciences sociales si on ne veut pas aller plus loin. Attention au nombre de variables qui ne respectent pas l'hypothèse, l'estimation de la baseline du risque pourrait être sensiblement affectée si l'analyse a des visées prédictives. Il convient tout de même lors de l'interprétation, de préciser les variables qui seront analysées sous cette forme très « moyenne » sur la période d'observation.

On peut également adapter cette stratégie du « ne rien faire » selon sens de l'altération des rapports de risque. Si aux cours du temps des écarts de risque, s'accroissent à la hausse comme à la baisse, on peut conserver cet estimateur moyen. La lecture des résultats ne sera pas affectée. Mais si cette non proportionnalité conduit à un changement du sens des rapport de risque je suis moins convaincu de la pertinence de cette stratégie.

Il faut également tenir compte de l'intérêt portée par les variables qui présentent un problème par rapport à l'hypothèse. Si on souhaite corriger la non proportionnalité avec une interaction, il est inutile de complexifier le modèle pour des variables introduites comme simples contrôles... qui ne sont là que pour jouer le rôle de contrôle. Et justement la vérification de la proportionnalité peut amener à s'interroger sur la bonne spécification du modèle.

En effet, on sait qu'une des causes du non respect de l'hypothèse peut provenir d'effets de sélection liées à des variables omises ou non observables. En analyse de durée ce problème prend le nom de **frailty** (fragilité) lorsque cette non homogénéité n'est pas observable. Des estimations, plus complexes, sont possibles dans ce cas, et sont en mesure malgré leur interprétation plutôt difficile de régler le problème. Il convient donc de bien spécifier le modèle au niveau des variables de contrôle observables et disponibles.

### *Modèle de Cox stratifié*

Utiliser la méthode dite de « Cox stratifiée » (non traitée). Utile si l'objectif est de présenter des fonctions de survie prédites ajustées, et si une seule covariable (binaire) présente un problème. Les HR ne seront pas estimés pour la variable qui ne respecte pas l'hypothèse: Le risque de base pour chaque groupe est alors estimé en amont.

### *Interaction*

Introduire une interaction avec la durée.

Cela permet d'enrichir le modèle au niveau de l'interprétation. Valable si peu de covariables présentent des problèmes d'uniformité dans le temps des rapports de risque, dans l'idéal une seule variable. Attention tout de même à la forme de la fonction, dans l'exemple on a contraint l'effet d'interaction à être strictement linéaire, cela facilite la lecture, mais c'est une hypothèse très forte.... car au final on introduit de nouveau une contrainte de proportionnalité dans le modèle.

### *Modèles alternatifs*

Utiliser un modèle alternatif: modèles paramétriques à risques proportionnels si la distribution du risque s'ajuste bien, le modèle paramétrique « flexible » de Parmar-Royston ou un modèle à durée discrète/groupée (voir section suivante). Pour la dernière solution, on peut également corriger la non

proportionnalité avec l'introduction d'une interaction. Si on ne le fait pas, les risques prédits qui sont par définition des probabilités conditionnelles, resteront toujours dans les bonnes bornes contrairement au modèle de Cox.

Utiliser un modèle non paramétrique additif dit d'Aalen ou une de ses variantes (non traité). Mais ces modèles, dont les résultats seront présentés par des graphiques, se commentent assez difficilement. Et dont l'utilisation est devenue je crois plus que rare.

### **Forêt aléatoire**

Autre méthode : les forêts aléatoires. L.Breiman a dès le départ proposé une estimation des modèles de survie par cette méthode. Par définition, pas sensible à l'hypothèse PH. Mais cela reste des méthodes à finalité très prédictive, moins riche en interprétation.



## 9 Modèle à durée discrète

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

- Par définition ce n'est pas un modèle à risques proportionnels, mais à **Odds proportionnels**. Toutefois en situation de rareté ( $p < 0.1$ ), l'Odds converge vers une probabilité, qui est une mesure du risque. Et donc les Odds ratio peuvent aisément se lire comme des rapports de probabilités.
- Le modèle à durée discrète est de type pleinement paramétrique, il est moins contraignant que le modèle de Cox si l'hypothèse de proportionnalité n'est pas respectée, car le modèle est directement ajusté par une fonction de la durée.
- Pour être estimé, la base de données doit être préalablement transformée en format long: sur les durées d'observation directement disponibles ou sur des regroupements de celles ci éventuellement de longueur différentes, ou sur les intervalles définis à partir des moments d'évènement lorsqu'il y a un nombre important d'évènements simultanés. (Kaplan-Meier et Cox).
- Ce modèle permet d'introduire bien plus simplement qu'avec le modèle de Cox un ensemble de covariables non fixes.

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

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

### **i** Les fonctions de lien

On restera ici sur la fonctions de lien logistique, mais bien évidemment les autres fonctions associées à la fonction de masse binomiale sont utilisables.

- Le lien probit, est a ma connaissance très peu utilisé en analyse des durées. C'est bien évidemment lié aux lectures des paramètres estimés. On peut utiliser des formes standardisés sur l'échelle des probabilités avec des effets marginaux, mais ceux-ci ne pourront pas être généralisés sur l'ajustement de la durée se fait avec une forme quadratique ou à la présence d'une interaction de quelque forme que ce soit (même problème si lien logistique).
- Le lien complémentaire log-log, est une alternative bien plus intéressante en raison de son lien avec la relation entre fonction de séjour et risque cumulé. Son utilisation doit être néanmoins d'être réservée à des situations de rareté des probabilités conditionnelles (disons  $< .10$ ,  $0.5$  préconisé), ce qui est souvent le cas en analyse de survie <sup>1</sup>. Sous cette condition, on peut interpréter directement les estimateurs comme des rapports de risque au sens de l'analyse de survie. On pourra se reporter à l'explication de G.Rodriguez dont le support est en lien la bibliographie du document. Je ferai peut-être un jour une section dédiée dans la partie annexe. La construction mathématique est plutôt simple.

- Pour information seulement car non testé. Développés dans le domaine l'épidémiologie, il existe des algorithmes forçant l'application du lien  $\log(p)$  à une vraisemblance de type binomiale. On parle de modèle *log-binomial*. Il s'agit bien de la fonction de lien canonique associée à la vraisemblance d'un processus poissonien<sup>2</sup> et dont les estimateurs sont indirectement interprétables en rapport de risques. Cette technique ne doit être envisagé qu'avec des incidences supérieures à 20%. Je n'ai pas d'expérience particulière sur ce type de modèles<sup>3</sup>, je ne suis donc pas capable de les évaluer en analyse de durée avec des durées très groupées, mais au niveau outils il semble que les solutions disponibles dans des packages R comme *logbin* soient à privilégier.

## 9.1 Organisation des données

### Format long

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

### Durée

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

### Variable évènement/censure

Si l'individu a connu l'évènement, elle prend la valeur 0 avant celui-ci. Au moment de l'évènement sa valeur est égale à 1. Pour les observations censurées, la variable prend toujours la valeur 0. **seule la dernière ligne observées peut prendre une valeur non nulle**<sup>4</sup>.

### Application

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

### Format d'origine

<sup>3</sup>la distribution des probabilités sous cette loi n'est pas, contrairement aux lois normale ou logistique, symétriques. Dans ce qui suit, il n'est pas conseillé de l'utiliser dans l'application avec l'ajustement sous forme d'indicateurs avec seulement 4 intervalles

<sup>3</sup>Non traité dans ce support, une méthode d'estimation en analyse de survie se nomme *counting process* et utilise des méthodes d'estimation de type poissonien

<sup>3</sup>je remercie néanmoins Emilie Counil et Nargès Gouhoubi pour l'information

<sup>4</sup>J'anticipe la version multinomiale pour les risques concurrents

Table 9.1: Durée discrète: données en format d'origine

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

La personne décède lors du deuxième intervalle de 30 jours

### Format long et variables pour l'analyse

Table 9.2: Durée discrète: données en format long

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

## 9.2 Ajustement de la durée

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

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

### **i** Le modèle de Cox à durée discrète/groupée

Cox est également à l'origine du modèle à durée discrète (1973). Par rapport aux pratiques courantes, la différence repose sur les bornes intervalles, identiques à celles définies pour l'estimation de la courbe de survie KM ou du modèle semi-paramétrique, à savoir une définition sur les moments d'évènement. Avec un ajustement de la durée reposant sur des indicatrices (ajustement sur des variables discrètes), ce modèle est quasiment identique au modèle semi paramétrique.

On peut remarquer qu'en sciences sociales, avec des durées assez fortement groupées, les intervalles directement observés et les intervalles définis aux moments d'évènements sont souvent identiques et s'ils ne le sont pas c'est souvent en tout début ou en toute fin de la période d'observation/exposition. En cas d'ajustement de la durée par des indicatrices, la définition des bornes des intervalles aux moments des évènements permet de s'assurer de la présence d'au moins une occurrence .

### 9.2.1 Ajustement avec une fonction quantitative de la durée

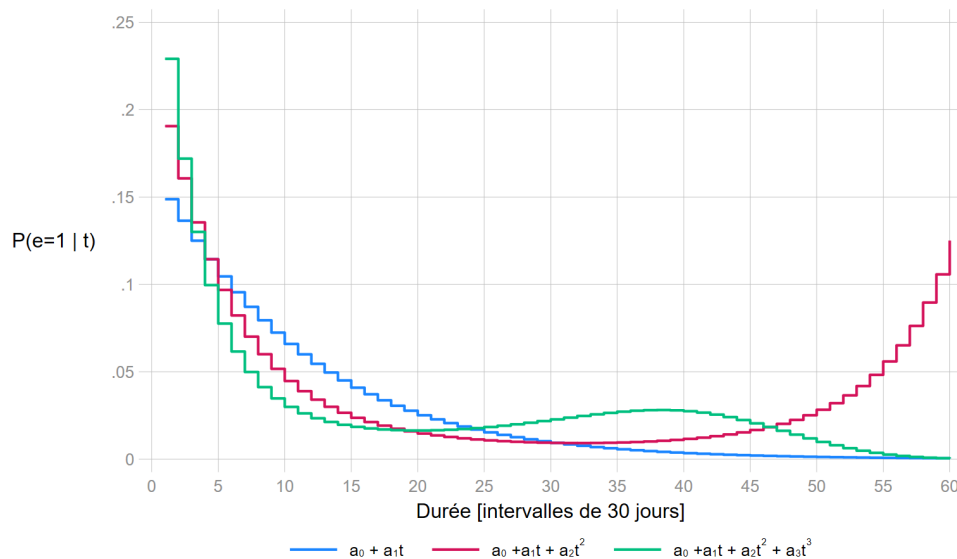
Le modèle étant paramétrique, on doit trouver une fonction qui ajuste le mieux les données. Toute transformation de la variable est possible:  $f(t) = a \times t$ ,  $f(t) = a \times \log(t)$ .....formes quadratiques. Les ajustements sous forme de **splines cubiques** tendent à se développer ces dernières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

Pour sélectionner cette fonction, on peut tester différents modèles sans covariable additionnelle [ $g(y(t)) = f(t)$ ], et sélectionner la forme dont le critère d'information de type vraisemblance pénalisée (**AIC**, **BIC**) est le plus faible, avec au moins des différences de -6 ou -8.

Exemple:

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

Figure 9.1: Probabilité de décéder avec 3 ajustements de la durée



## Critères AIC

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

On peut utiliser la troisième forme à savoir  $a_1 \times t + a_2 \times t^2 + a_3 \times t^3$  <sup>5</sup>.

## Estimation du modèle avec toutes les covariables

Table 9.4: Modèle logistique à durée discrète ( $f(t)$  continue)

Variables	OR - RR	Std. err	z	P> z	95% IC
$t$	0.678	0.057	-4.52	0.000	0.587 ; 0.810
$t^2$	1.014	0.005	+2.83	0.005	1.004 ; 1.024

<sup>5</sup>Ce n'est pas le cas ici, mais si on sélectionne une forme cubique, je conseille vivement de regarder les probabilités conditionnelles prédites, en particulier en fin de périodes d'observation/exposition si peu d'individus restent soumis au risque. On peut rencontrer des problèmes d'overfitting avec des probabilités conditionnelles estimées trop proche de 1. Pour les personnes qui suivent la formation, c'est le cas avec les données des TP

Variables	OR - RR	Std. err	z	P> z	95% IC
$t^3$	1.000	0.000	-2.11	0.035	1.000 ; 1.000
<i>year</i>	0.876	0.015	-1.80	0.072	0.758 ; 1.012
<i>age</i>	1.034	0.163	+2.27	0.023	1.005 ; 1.064
<i>surgery</i>	0.364	0.110	-2.25	0.024	0.151 ; 0.877
<i>Constante</i>	<i>0.440</i>	<i>0.110</i>	<i>-3.29</i>	<i>0.001</i>	<i>0.270 ; 0.718</i>

*Remarque:* les variables *year* et *age* ont été centrées sur leur moyenne pour rendre la constante interprétable. La constante reporte donc l'Odds de décéder lors des 30 premiers jours d'une personne dont l'âge et l'année à l'entrée dans le registre à l'âge moyen et à l'année moyenne, et qui n'a pas été opéré préalablement pour un pontage.

Si maintenant on estime un modèle de Cox sur ces données journalières groupées, on remarque que les résultats obtenus, et ce n'est pas une surprise, sont très proches.

Table 9.5: Modèle de Cox

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

## 9.2.2 Ajustement discret

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

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

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

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

Table 9.6: Modèle logistique à durée discrète ( $f(t)$  indicatrices)

Variables	OR	Std. err	z	P> z	95% IC
0 – 4 <i>mois</i>	2.811	1.177	+2.47	0.014	1.237 ; 6.387
5 – 11 <i>mois</i>	ref	-	-	-	-
12 – 23 <i>mois</i>	0.559	0.346	-0.94	0.347	0.166 ; 1.881
24 – 46 <i>mois</i>	1.741	1.159	+0.83	0.405	0.472 ; 6.417
<i>year</i>	0.816	0.076	-2.18	0.029	0.680 ; 0.980
<i>age</i>	1.048	0.019	+2.53	0.011	1.011 ; 1.087
<i>surgery</i>	0.330	0.166	-2.21	0.027	0.123 ; 0.882
<i>Constante</i>	0.407	0.151	2.43	0.015	0.198 ; 0.840

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

Variables	Ajustement discret	Ajustement continu
<i>year</i>	0.816	0.876
<i>age</i>	1.048	1.034
<i>surgery</i>	0.330	0.364

### **i** Lien avec des modèles usuels à durée continue

Si la durée discrète/groupée sous tend une durée continue (ce qui est clairement le cas ici):

- L'ajustement avec des durées sous forme d'indicatrices correspond au modèle à durée discrète défini par Cox. Il est également assimilable à un modèle appelé *exponential piecewise constant* (soit un modèle de poisson).
- Si l'ajustement se fait en utilisation une transformation de la durée par une fonction:
  - $f(t) = \log(t)$  correspond à un modèle de Weibull à *risque proportionnel* <sup>6</sup>.
- Si l'ajustement se fait avec des splines cubiques, le modèle à durée discrète correspond à un modèle de type *Parmar-Royston*.

## 9.3 Proportionnalité des risques

- Formellement un modèle logistique à durée discrète/groupée repose sur une hypothèse d'Odds proportionnels [Odds ratios constants pendant la durée d'observation]. Contrairement au modèle de Cox, l'estimation des probabilités (risque) n'est pas biaisée si l'hypothèse PH n'est pas respectée, les paramètres estimés seront considérés au pire comme des approximations.... problématique d'interaction oblige.

<sup>6</sup>Voir la très courte section sur les modèles paramétriques usuels

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

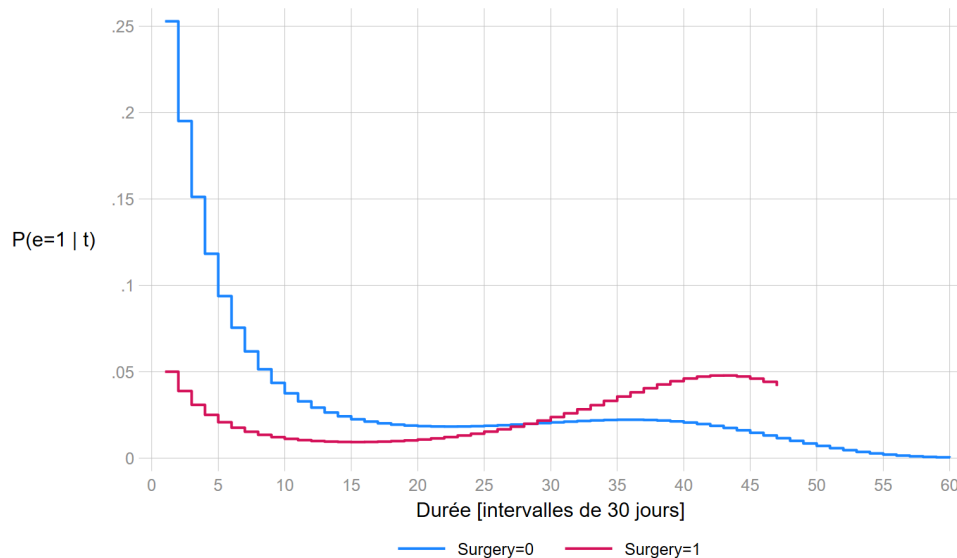
Avec un ajustement continue, on remarque de nouveau que le résultat du modèle est très proche de celui estimé avec un modèle de Cox.

Table 9.8: Modèle logistique à durée discrète avec correction de la non proportionnalité

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

Avec une spécification dépendant seulement de la durée (pour le graphique:

Figure 9.2: Probabilité de décéder après correction de la non proportionnalité pour la variable surgery



# 10 Variables dynamiques

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

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

## 10.1 Facteur dynamique traitée de manière fixe

On reprend l'exemple sur malformation cardiaque, en ajoutant la variable relative à la greffe. La question est donc de savoir si une transplantation du coeur réduit le risque journalier de décéder (ou augmente la durée journalière de survie).

On a dans la base 2 variables: une variable binaire pour savoir si l'individu à été greffé ou non, **transplant**, et la variable *wait* de type continue tronquée donnant la durée en jour jusqu'à l'opération depuis l'inscription dans le registre (0 si *transplant* = 0).

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

Table 10.1: Modèle de cox avec une variable dynamique (binaire) traitée de manière fixe (estimation biaisée)

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

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



Sauf que.....

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

Table 10.2: Mapping de la base avec une variable dynamique binaire traitée de manière fixe

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

Une personne est codée greffée avant le jour de la transplantation. Si on cherche à mesurer un **effet causal**, il est donc mal spécifié. Sa dimension temporelle ayant été ignorée, soit ici le jour exact de l'opération. C'est le même principe pour l'évènement, la personne est codée décédée (1) le jour du décès, et vivante avant (0).

## 10.2 Estimation avec une variable dynamique

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

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

### 10.2.1 Modèle de Cox

Table 10.3: Mapping correct de la base avec une variable dynamique binaire

id	year	age	surgery	transplant	wait	died	$t_0$	$t$	<b>TVC</b>
70	72	52	0	1	5	0	0	1	<b>0</b>
70	72	52	0	1	5	0	1	2	<b>0</b>
70	72	52	0	1	5	0	2	3	<b>0</b>
70	72	52	0	1	5	0	3	5	<b>0</b>

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

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

Table 10.4: Modèle de Cox avec une variable dynamique binaire

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

L'impact de la greffe apparaît maintenant bien plus modéré sur la survie des individus. Cela ne signifie pas non plus que des personnes ont pu être *sauvée* grâce à cette opération (ou plutôt leur durée de vie augmentée), mais des complications lors de l'opération ou post-opératoire, surtout à une époque où ces techniques étaient à leurs balbutiements, ont pu également accélérer la mortalité. Il faut également garder en tête que l'état de santé des personnes est particulièrement dégradé, cette opération étant celle de la *dernière chance*.

R - Stata - Python - Sas

### 10.2.1.1 R - Stata, Python

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

### 10.2.1.2 Sas †

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

...

## 10.2.2 Modèle à temps discret

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

Table 10.5: Modèle logistique à durée discrète avec variable dynamique binaire

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

## 10.3 Précautions à prendre

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

**partie V**

**Compléments**

# 11 Éléments de mise en forme des données

Ce qui suit est un premier draft réalisé en 2023, encore très perfectible, mais j'avais conscience depuis plusieurs années que l'absence d'une partie dédiée aux manipulations des données biographiques était problématique. Ce qui suit ne peut pas couvrir toutes les problématiques que l'on peut rencontrer lorsqu'on met en forme ce type de données: le format de mise à disposition, l'extraction de l'information pertinente correspondant à la question analysée, voire le logiciel utilisé, rend quasiment impossible la production d'un guide clé en main.

## Quelques conseils

- Comme l'analyse des durées/survie consiste à extraire une (ou plusieurs) séquence(s), il convient de bien identifier les situations types que l'on sera conduit à rencontrer et à traiter. Ici il peut être intéressant, de manière exploratoire, de conduire une analyse de séquences au préalable.
- Vérifier à chaque étape de la programmation la validité ou la cohérence des informations contenues dans les variables créées et/ou de la mise en forme.

## Le format des données utilisé dans cette section

- Il est identique à celui utilisé dans l'enquête MAFE [\[lien\]](#)<sup>1</sup>, donc de type individus-séquences avec un marquage des informations temporelles de type *année*.
- Sur chaque séquence l'année de début et l'année de fin sont indiquée.
- Les trajectoires peuvent être continues ou non. Si elle ne sont pas continues, un exemple montre comment récupérer les séquences en "creux".

## Programmation

- Cette section n'est pas centrée sur la méthode de programmation, mais plutôt sur les questions que l'on doit se poser et leur résolution.
- Comme la formation est appliquée en R, j'ai néanmoins donné les codes que j'ai utilisés.
  - Les codes R sont ici largement hérité d'une programmation Stata que je maîtrise bien mieux, et particulièrement efficace dans ce domaine.
  - On pourrait aller *plus vite*, mais j'assume une programmation prudente.
  - Le type de programmation applique assez massivement l'utilisation de compteurs ou de double compteurs, et si nécessaire l'utilisation de variables décalée de type *lead* ou *lag*.

### **i** Le support **assistooldsms**

Le service SMS de l'Ined a mis en place récemment un support de programmation dédié principalement à R [\[Lien\]](#). Ce support est construit sur l'idée d'une liste de fiches thématiques qui trouvent

<sup>1</sup>Des éléments de manipulation/programmation pour un exemple volontairement très compliqué sont donnés dans *méthodes => notes méthodologiques*. Ayant été fait en 2015, le code pour R est largement *out of date*

leur origine, mais pas que, dans des demandes d’assistances de programmation de la part de chercheur.e.s, (post)doctorant.e.s ou stagiaires. Bien que cela ne soit pas pour tout de suite, il est prévu d’alimenter ce support sur la question des manipulations des données biographiques avec: d’autres exemples, des alternatives aux codes proposés plus loin, ou l’application des exemples traités ici mais avec un format de mise à disposition des données différents.  
... Ceci au mieux pour l’été 2024.

Packages utilisés:

```
library(dplyr)
library(tidyr)
library(knitr)
```

## 11.1 Calcul des variables d’analyses

On partira de la base individus-séquences suivante:

```
df = data.frame(id = c(1, 1, 1, 2),
                deb = c(2020, 2023, 2024, 2022),
                fin = c(2021, 2024, 2025, NA),
                x = c(1,2,1,2))
kable(df)
```

id	deb	fin	x
1	2020	2021	1
1	2023	2024	2
1	2024	2025	1
2	2022	NA	2

On supposera que l’année de collecte, pour toutes les observations, est **2025** <sup>2</sup>.

Si cela n’est pas donné dans le module biographique, il peut être intéressant de construire les numéros de séquences des trajectoires.

```
df$nseq = 1
df = df %>% group_by(id) %>% mutate(nseq = cumsum(nseq))
kable(df)
```

<sup>2</sup>Ici on a une enquête réalisée une même année pour toute les observations, ce n’est pas toujours le cas. De même au lieu de l’année, si les datations avaient été données par l’âge, au moment de l’enquête l’âge varierait d’une personne à une autre. Ces datations différentes (année ou âge) peuvent être présentes dans chaque module biographique d’une enquête, ou dans le fichier des caractéristiques fixes. Dans ce cas l’information devra être récupérée

id	deb	fin	x	nseq
1	2020	2021	1	1
1	2023	2024	2	2
1	2024	2025	1	3
2	2022	NA	2	1

### Exemple 1 : durée de séjour de la première séquence observée

Supposons que *x* traduit un type de relation/union, par exemple *x*=1 est une relation non cohabitante et *x*=2 est une relation cohabitante. On s'intéresse à la durée de la première relation, sans distinction entre 1 et 2. Il suffit de sélectionner la première séquence.

```
df = filter(df, nseq==1)
```

La variable de *fin* va permettre de repérer les informations censurées, et de générer la variable d'évènement. A ce niveau il est donc important de ne pas encore remplacer la date de censure par sa valeur.

- Si *fin* est une valeur manquante: observation censurée.
- Si *fin* est une valeur renseignée: occurrence de l'évènement.

```
df$e = ifelse(is.na(df$fin), 0,1)
kable(df)
```

id	deb	fin	x	nseq	e
1	2020	2021	1	1	1
2	2022	NA	2	1	0

Pour la variable de durée <sup>3</sup>, une repéré les observations censurées, elle est calculée directement avec les variables *fin* et *deb*.

```
df$dur = ifelse(df$e==1, df$fin - df$deb + 1, 2025 - df$deb + 1)
kable(df)
```

<sup>3</sup>La mesure est ici discrète/groupée, il me semble toujours préférable d'allonger les durées à +1. On démarre donc toujours un premier janvier pour terminer un 31 décembre sur l'information est donnée par des année. Ici *t*=1 représente la première année après la sortie des études. Une personne qui aura eu un emploi durant cette année, l'aura eu durant cette première année, que ce soit 2 semaines après ou 11 mois après. Si on disposait des mois, cela pourrait être intéressant de modifier cette métrique temporelle. Voir exemple 3

id	deb	fin	x	nseq	e	dur
1	2020	2021	1	1	1	2
2	2022	NA	2	1	0	4

## Exemple 2 : changement de métrique temporelle

Toujours avec le même exemple, mais en ajoutant une observation, supposons que l'on dispose également de l'information sur les mois. Sur les mois où l'évènement à eu lieu, mais également sur les mois où l'enquête a été réalisée.

```
df2 = data.frame(id = c(1, 1, 1, 2,3),
                 deb = c(2020, 2023, 2024, 2022, 2021),
                 debm = c(2,5,3,10,9),
                 fin = c(2021, 2024, 2025, NA,2021),
                 finm = c(4,2,12,NA,11),
                 x = c(1,2,1,2,1),
                 enq = c(2025,2025,2025,2025,2025),
                 enqm = c(4,4,4,5,4))

df2$nseq = 1
df2 = df2 %>% group_by(id) %>% mutate(nseq = cumsum(nseq))

kable(df2)
```

id	deb	debm	fin	finm	x	enq	enqm	nseq
1	2020	2	2021	4	1	2025	4	1
1	2023	5	2024	2	2	2025	4	2
1	2024	3	2025	12	1	2025	4	3
2	2022	10	NA	NA	2	2025	5	1
3	2021	9	2021	11	1	2025	4	1

On remarque que la nouvelle observation (id=3) a connu l'évènement, ici la fin de la relation, la même année qu'au début d'exposition (le début de la relation)... mais au bout de 2,6,11 mois???? Comme on dispose de l'information sur les mois de début et de fin cela peut être intéressant de l'exploiter. De la même manière si l'enquête a été réalisée la même année, les entretiens n'ont pas eu lieu le même mois. On aura besoin de cette information pour les observations censurées.

De nouveau on sélectionne la première séquence, et pour la lisibilité de la base on retire les informations qui ne seront pas ou plus exploitées (*nseq*, *x*).

```
df2 = filter(df2,nseq==1)
df2 = select(df2, -c(x,nseq))

kable(df2)
```



id	deb	debm	fin	finm	enq	enqm
1	2020	2	2021	4	2025	4
2	2022	10	NA	NA	2025	5
3	2021	9	2021	11	2025	4

On génère la variable censure/événement (toujours à faire avant la variable de durée) de la même manière que pour l'exemple 1.

```
df2$e = ifelse(is.na(df2$fin), 0, 1)
kable(df2)
```

id	deb	debm	fin	finm	enq	enqm	e
1	2020	2	2021	4	2025	4	1
2	2022	10	NA	NA	2025	5	0
3	2021	9	2021	11	2025	4	1

Pour la variable de durée, le principe est de multiplier par 12 la différence entre l'année de fin et l'année de début et d'ajouter la différence entre le mois de fin et le mois de début.

Pour les observations censurées, ici l'année de fin est identique mais les mois varient. En terme de programmation, surtout si avec R on utilise `ifelse`, il est préférable d'y aller doucement en créant une durée pour les observations qui ont connu l'évènement et une durée pour les observations censurées. Puis de regrouper les deux cas. C'est ce qui est fait dans le code qui suit.

Durée selon les valeurs de *e*:

```
df2$dur1 = ifelse(df2$e==1, 12*(df2$fin - df2$deb) + (df2$finm - df2$debm), 0)
df2$dur0 = ifelse(df2$e==0, 12*(2025 - df2$deb) + (df2$enqm - df2$debm), 0)
kable(df2)
```

id	deb	debm	fin	finm	enq	enqm	e	dur1	dur0
1	2020	2	2021	4	2025	4	1	14	0
2	2022	10	NA	NA	2025	5	0	0	31
3	2021	9	2021	11	2025	4	1	2	0

On regroupe par simple sommation (le *else* étant 0).

```
df2$dur = df2$dur1 + df2$dur0
df2 = select(df2, -c(dur1,dur0))
kable(df2)
```

id	deb	debm	fin	finm	enq	enqm	e	dur
1	2020	2	2021	4	2025	4	1	14
2	2022	10	NA	NA	2025	5	0	31
3	2021	9	2021	11	2025	4	1	2

On dispose ainsi des éléments nécessaires pour faire une analyse de durée avec une métrique mensuelle <sup>4</sup>.

### Exemple 3 : importation d'un début d'exposition externe

On repart de la première base

id	deb	fin	x	nseq
1	2020	2021	1	1
1	2023	2024	2	2
1	2024	2025	1	3
2	2022	NA	2	1

On suppose maintenant que  $x$  traduit des situations sur le marché du travail. Par exemple  $x=1$  est un emploi en CDD et  $x=2$  un emploi en CDI. On s'intéresse à la durée entre la fin des études et le premier emploi, quel que soit son type.

- On ne dispose pas ici de toutes les informations pour calculer la durée, soit la fin des études. Elle peut être donnée dans une base classique regroupant l'ensemble des caractéristiques individuelles de type fixe (année de naissance, sexe...).
- Comme on s'intéresse à la durée de recherche du premier emploi, dans le module biographique la date de début va devenir la date de fin.
- Pour les observations présentes dans la base biographique, il n'y a pas de censure à droite. Mais si on regarde le fichier des caractéristiques générales, fixe:

```
etude = data.frame(id = c(1,2,3), fin_etude = c(2020,2021,2023))
kable(etude)
```

id	fin_etude
1	2020
2	2021
3	2023

Une nouvelle observation ( $id=3$ ) apparaît. Au moment de l'enquête, elle n'a pas (**encore**) trouvé un emploi depuis la fin de ces études. On a donc une observation qui sera censurée.

<sup>4</sup>Contrairement à la durée annuelle je n'ai pas ajouté 1 à chaque durée, ce qui est de nouveau envisageable par exemple si on veut explicitement indiquer les événements qui ont lieu le premier mois. Pour  $id=3$  la relation a-t-elle durée du 1er septembre au 30 novembre, ou du 30 septembre au 1er novembre?? On a toujours un problème de précision, mais ici d'une trentaine de jours

### Note

Certaines bases biographiques peuvent être structurées avec des trajectoires strictement continue, l'année (l'âge) de fin étant l'année (l'âge) de début de la trajectoire suivante. Dans ce cas, l'information serait immédiatement disponible, avec la présence d'un nombre de séquences plus important dans la base.

On va devoir:

- Sélectionner la première séquence d'emploi dans la base *df* (variable *nseq*).
- La fusionner avec la base étude.

Avant la fusion, on peut conserver seulement les informations nécessaires (*id*, *deb*). La variable *deb* va changer également de statut en devenant l'année de *fin* de la période de recherche d'emploi.

```
df = filter(df, nseq==1)
df = select(df, -c(fin,x,nseq))

df = rename(df, fin = deb)
kable(df)
```

id	fin
1	2020
2	2022

Après la fusion:

```
df = full_join(etude, df, by = c('id'))

df = rename(df, deb = fin_etude)

kable(df)
```

id	deb	fin
1	2020	2020
2	2021	2022
3	2023	NA

On a toutes les informations pour générer la variable censure/événement et la variable de durée:

```
df$e = ifelse(is.na(df$fin),0,1)

df$dur = ifelse(df$e, df$fin - df$deb + 1, 2025 - df$deb + 1)
kable(df)
```

id	deb	fin	e	dur
1	2020	2020	1	1
2	2021	2022	1	2
3	2023	NA	0	3

## 11.2 Appariement de modules biographiques

On repart de la première base, avec les numéros de séquence.

id	deb	fin	x	nseq
1	2020	2021	1	1
1	2023	2024	2	2
1	2024	2025	1	3
2	2022	NA	2	1

### 11.2.1 Mise en forme d'une base

Pour appairer des informations de plusieurs modules biographiques, on doit transformer les bases en format individus-séquences en format individus-périodes (ici individus années).

- **Etape 1:** allongement sur chaque séquence après avoir générées leur durée
- **Etape 2:** générer une variable de période (année) sur chaque ligne. Elle servira pour l'appariement.

**Pourquoi ne pas utiliser la simple différence entre la fin et le début ?**

Durée (fin - début) et allongement de la base:

On ne génère pas des variables d'analyse, on aurait besoin de l'information sur l'année de l'enquête pour les informations censurées.

```
df$fin[is.na(df$fin)] = 2025
kable(df)
```

id	deb	fin	x	nseq
1	2020	2021	1	1
1	2023	2024	2	2
1	2024	2025	1	3
2	2022	2025	2	1

Allongement de la base:

```
df1 = df
df1$dur1 = df1$fin - df1$deb

df1$dur1b = df1$dur1 # uncount supprime la variable d'origine
df1 = uncount(df1,dur1b)

kable(df1)
```

	id	deb	fin	x	nseq	dur1
	1	2020	2021	1	1	1
	1	2023	2024	2	2	1
	1	2024	2025	1	3	1
	2	2022	2025	2	1	3
	2	2022	2025	2	1	3
	2	2022	2025	2	1	3

Pour générer la variable période (année), on a besoin d'un compteur qui sera associé à la variable *deb*. On doit bien contrôler l'opération par identifiant et numéro de séquence.

```
df1$c = 1
df1 = df1 %>% group_by(id,nseq) %>% mutate(year = deb + cumsum(c))

kable(df1)
```

	id	deb	fin	x	nseq	dur1	c	year
	1	2020	2021	1	1	1	1	2021
	1	2023	2024	2	2	1	1	2024
	1	2024	2025	1	3	1	1	2025
	2	2022	2025	2	1	3	1	2023
	2	2022	2025	2	1	3	1	2024
	2	2022	2025	2	1	3	1	2025

**Problème:** les années de début ne sont pas correctes: 2021 au lieu de 2020 pour la première séquence de id=1 par exemple.

### ! Important

On doit donc impérativement augmenter la différence entre la fin et le début par +1 pour que l'ensemble des périodes (années) soit couvertes.

On reprend donc les opérations précédentes mais avec  **$\text{durée} = \text{fin} - \text{debut} + 1$**

- Allongement de la base avec durée augmentée

```
df2 = df
df2$dur2 = df2$fin - df2$deb + 1

df2$dur2b = df2$dur2 # uncount supprime la variable d'origine
df2 = uncount(df2,dur2b)

kable(df2)
```

id	deb	fin	x	nseq	dur2
1	2020	2021	1	1	2
1	2020	2021	1	1	2
1	2023	2024	2	2	2
1	2023	2024	2	2	2
1	2024	2025	1	3	2
1	2024	2025	1	3	2
2	2022	2025	2	1	4
2	2022	2025	2	1	4
2	2022	2025	2	1	4
2	2022	2025	2	1	4

- Création de la variable *year*: sur chaque individus-séquences, la somme entre le compteur et l'année de début doit être réduite de 11.

```
df2$c = 1
df2 = df2 %>% group_by(id,nseq) %>% mutate(year = deb + cumsum(c) - 1)

df2 = select(df2, -c(deb,fin,dur2))

kable(df2)
```

id	x	nseq	c	year
1	1	1	1	2020
1	1	1	1	2021
1	2	2	1	2023
1	2	2	1	2024
1	1	3	1	2024
1	1	3	1	2025
2	2	1	1	2022
2	2	1	1	2023
2	2	1	1	2024
2	2	1	1	2025

Les années sont toutes couvertes....mais un peu trop. En effet, lorsque les trajectoires sont continues soit lorsque l'année de fin d'une séquence est identique à l'année de début de la suivante, les années vont être doublonnées. On doit donc supprimer ce doublon.

- Suppression des doublons des trajectoires continues.

De nouveaux on doit faire un choix, soit on privilégie l'année de fin, soit on privilégie l'année de début. Les applications ont des fonctions qui permettent de supprimer les doublons<sup>5</sup>. On peut le faire manuellement en regardant pour chaque personnes-années le nombre de doublon. Cela se fait facilement à l'aide d'un compteur, ici la variable *nyear*.

```
df2 = df2 %>% group_by(id,year) %>% mutate(nyear = cumsum(c))
kable(df2)
```

id	x	nseq	c	year	nyear
1	1	1	1	2020	1
1	1	1	1	2021	1
1	2	2	1	2023	1
1	2	2	1	2024	1
1	1	3	1	2024	2
1	1	3	1	2025	1
2	2	1	1	2022	1
2	2	1	1	2023	1
2	2	1	1	2024	1
2	2	1	1	2025	1

Si on souhaite garder l'année de fin on filtre les observations en conservant celles dont *nyear*=1. Si on souhaite privilégier les années de début on filtre les observations en conservant celles dont *nyear*=2. Si on souhaite conserver les années de fin de séquence:

```
df2 = filter(df2, nyear==1)
df2 = select(df2, -c(nseq,c,nyear))
kable(df2)
```

id	x	year
1	1	2020
1	1	2021
1	2	2023
1	2	2024

<sup>5</sup>avec R par exemple la fonction *unique* de *dplyr*

id	x	year
1	1	2025
2	2	2022
2	2	2023
2	2	2024
2	2	2025

### ! En résumé

- A la date (année/âge) de censure remplacer la valeur manquante par sa valeur. Si ultérieurement on a besoin de garder l'information sur la censure - valeur manquante -, on peut générer une variable miroir de *fin*.
- Sur chaque séquence calculer la durée avec une augmentation de +1.
- Créer une variable période (année) sur chaque ligne. Elle servira à définir la clé d'appariement.
- Supprimer les doublons sur les transition continue  $fin_t = debut_{t+1}$ .

## 11.2.2 Fusion des informations biographiques

### 11.2.2.1 Fusion avec l'ensemble des périodes observables

Pour commencer par un exemple plutôt simple, on note que pour id=1 l'année 2022 n'est pas renseignée (trajectoire non continue). Si on reprend un exemple précédent (relations de couple), cette année pourrait être identifiée comme une période sans relation. Une façon simple de boucher ce type "trous", est d'utiliser les années de naissances des individus, et de créer une base individus-périodes qui couvre toutes les années de vie de l'individu jusqu'à l'enquête. On remontera jusque là, mais on va par exemple considérer que pour id=1 et id=2 ce *début de tout* est en 2018.

```
dftout = data.frame(id = c(1, 2),
                    t0 = c(2018, 2018))

kable(dftout)
```

id	t0
1	2018
2	2018

- On ajoute l'information sur l'année de l'enquête (2025).
- On génère la durée
- On allonge la base
- On génère la variable année sur chaque ligne (on contrôle seulement sur *id*)



```

dftout$tmax = 2025

dftout$dur = dftout$tmax - dftout$t0 + 1

dftout = uncount(dftout,dur)

dftout$c = 1
dftout = dftout %>% group_by(id) %>% mutate(year = t0 + cumsum(c) - 1)

dftout = select(dftout, -c(t0,tmax,c))

kable(dftout)

```

id	year
1	2018
1	2019
1	2020
1	2021
1	2022
1	2023
1	2024
1	2025
2	2018
2	2019
2	2020
2	2021
2	2022
2	2023
2	2024
2	2025

On peut maintenant apparier cette couverture de toutes les années de vie jusqu'à l'enquête à la base biographique:

```

df2 = full_join(df2, dftout, by = c("id","year"))

df2 = arrange(df2, id, year)
kable(df2)

```

id	x	year
1	NA	2018
1	NA	2019

id	x	year
1	1	2020
1	1	2021
1	NA	2022
1	2	2023
1	2	2024
1	1	2025
2	NA	2018
2	NA	2019
2	NA	2020
2	NA	2021
2	2	2022
2	2	2023
2	2	2024
2	2	2025

Pour supprimer les informations qui précèdent la première séquence de la biographie, on peut générer un compteur sur la variable x après avoir remplacé ses valeurs manquantes par des 0. On gardera les lignes pour lesquels ce compteur est supérieur à 1.

```
df2$x[is.na(df2$x)] = 0

df2 = df2 %>% group_by(id) %>% mutate(nx = cumsum(x))

kable(df2)
```

id	x	year	nx
1	0	2018	0
1	0	2019	0
1	1	2020	1
1	1	2021	2
1	0	2022	2
1	2	2023	4
1	2	2024	6
1	1	2025	7
2	0	2018	0
2	0	2019	0
2	0	2020	0
2	0	2021	0
2	2	2022	2
2	2	2023	4
2	2	2024	6
2	2	2025	8

On supprime les lignes lorsque nx=0.

```
df2 = filter(df2, nx>0)

df2 = select(df2, -c(nx))

kable(df2)
```

id	x	year
1	1	2020
1	1	2021
1	0	2022
1	2	2023
1	2	2024
1	1	2025
2	2	2022
2	2	2023
2	2	2024
2	2	2025

### 11.2.2.2 Fusion avec une autre base biographique

On peut être amené à fusionner plusieurs modules biographique. Jusqu'à présent, une même année, tous les individus ne pouvaient être que dans une situation, par exemple un seul emploi, un seul lieu de résidence etc... Pour certains phénomènes, une même années ou pendant une période plus longue on peut observer simultanément plusieurs états différent, ou plus classiquement observer une somme d'un même état. On parle ici d'*overlapping*. Ce type de situation est typiquement celle qu'on observe avec le nombre d'enfants.

Supposons que le base ci-dessous traduit la naissance et potentiellement le décès des enfants.

```
dfy = data.frame(id = c(1, 2, 2),
  deb = c(2022, 2019, 2023),
  fin = c(NA, 2024, NA),
  nseq = c(1,1,2))

kable(dfy)
```

id	deb	fin	nseq
1	2022	NA	1
2	2019	2024	1
2	2023	NA	2

- id=1 a un premier enfant en 2022 qui est toujours en vie au moment de l'enquête (2025)

- id=2:
  - A un premier enfant en 2019 qui décède en 2024
  - A un second enfant en 2023, toujours en vie au moment de l'enquête
  - De la naissance du second enfant au décès du premier, on va donc avoir des doublons (overlapping) sur les années

Si on reprend les manipulations précédentes jusqu'à la création de la variable *year*:

```
dfy$fin[is.na(dfy$fin)] = 2025
dfy$dur = dfy$fin - dfy$deb + 1

dfy$durb = dfy$dur # Uncount supprime la variable d'origine
dfy = uncount(dfy,durb)

dfy$c = 1
dfy = dfy %>% group_by(id,nseq) %>% mutate(year = deb + cumsum(c) - 1)
```

La variable *year* est bien renseignée 2 fois pour les années 2023 et 2024.

On peut s'intéresser au fait d'avoir ou non un enfant, ou de manière plus générale au nombre d'enfant. En créant cette information, on se donne également le moyen de corriger cet overlapping:

- On peut de nouveau générer un compteur contrôlé par individu année
- En générant un total de ligne doublonnée, on récupérera par exemple ici le nombre d'enfant en vie chaque année.
- En ne gardant que la ligne ou le compteur est égal à 1, on supprime les doublons tout en gardant l'information sur le nombre d'enfant en vie une année donnée.

```
dfy = dfy %>% group_by(id,year) %>% mutate(ny = cumsum(c))
dfy = dfy %>% group_by(id,year) %>% mutate(tot_y = sum(c))

kable(dfy)
```

id	deb	fin	nseq	dur	c	year	ny	tot_y
1	2022	2025	1	4	1	2022	1	1
1	2022	2025	1	4	1	2023	1	1
1	2022	2025	1	4	1	2024	1	1
1	2022	2025	1	4	1	2025	1	1
2	2019	2024	1	6	1	2019	1	1
2	2019	2024	1	6	1	2020	1	1
2	2019	2024	1	6	1	2021	1	1
2	2019	2024	1	6	1	2022	1	1
2	2019	2024	1	6	1	2023	1	2
2	2019	2024	1	6	1	2024	1	2
2	2023	2025	2	3	1	2023	2	2

id	deb	fin	nseq	dur	c	year	ny	tot_y
2	2023	2025	2	3	1	2024	2	2
2	2023	2025	2	3	1	2025	1	1

Il ne reste plus qu'à supprimer les lignes où  $ny > 1$

```
dfy = filter(dfy, ny==1)
dfy = select(dfy, -c(ny,deb, fin, dur, nseq, c))

kable(dfy)
```

id	year	tot_y
1	2022	1
1	2023	1
1	2024	1
1	2025	1
2	2019	1
2	2020	1
2	2021	1
2	2022	1
2	2023	2
2	2024	2
2	2025	1

Avec une ligne par année, on peut la fusionner avec une autre base biographique en format individus-années (même principe qu'avec la fusion avec la base sur toutes les années de vie).

```
df2y = full_join(dfy, df2, by = c("id","year"))

df2y = arrange(df2y, id,year)

df2y = select(df2y, c(id,year,x,tot_y))

df2y$tot_y[is.na(df2y$tot_y)] = 0
df2y$x[is.na(df2y$x)] = 0

kable(df2y)
```

id	year	x	tot_y
1	2020	1	0
1	2021	1	0
1	2022	0	1

id	year	x	tot_y
1	2023	2	1
1	2024	2	1
1	2025	1	1
2	2019	0	1
2	2020	0	1
2	2021	0	1
2	2022	2	1
2	2023	2	2
2	2024	2	2
2	2025	2	1

## 11.3 Sélection d'un type de séquence et mise en forme pour l'analyse

## 11.4 Durée jusqu'à la première séquence

```
df = data.frame(id = c( 1, 1, 1, 2, 3, 3, 4),
  deb = c(2018, 2022, 2024, 2019, 2023, 2024, 2023),
  fin = c(2021, 2024, 2025, NA, 2024, NA, NA),
  y = c(1, 2, 1, 2, 3, 2, 1),
  nseq = c(1, 2, 3, 1, 1, 2, 1)
)

kable(df)
```

id	deb	fin	y	nseq
1	2018	2021	1	1
1	2022	2024	2	2
1	2024	2025	1	3
2	2019	NA	2	1
3	2023	2024	3	1
3	2024	NA	2	2
4	2023	NA	1	1

On va s'intéresser à la durée jusqu'à l'occurrence de la séquence de type 2 ou 3 (variable *y*). On considérera que le début de l'exposition est donné par la variable *deb* sur la première séquence.

- id=1: début de l'exposition/observation en 2018, observe l'évènement en 2022.
- id=2: début de l'exposition/observation en 2019, observe l'évènement la même année.
- id=3: début de l'exposition/observation en 2019, observe l'évènement la même année.

- id=4: début de l'exposition/observation en 2023, n'a pas connu l'évènement au moment de l'enquête.

## Recupération de l'année de l'évènement

On peut repérer la présence d'une des deux séquences d'intérêt avec une indicatrice.

```
df$e = ifelse(df$y==2 | df$y==3,1,0)

kable(df)
```

id	deb	fin	y	nseq	e
1	2018	2021	1	1	0
1	2022	2024	2	2	1
1	2024	2025	1	3	0
2	2019	NA	2	1	1
3	2023	2024	3	1	1
3	2024	NA	2	2	1
4	2023	NA	1	1	0

De nouveau l'utilisation d'un compteur sur cette variable indicatrice, peut s'avérer utile pour repérer le moment de l'occurrence.

```
df = df %>% group_by(id) %>% mutate(n = cumsum(e))

kable(df)
```

id	deb	fin	y	nseq	e	n
1	2018	2021	1	1	0	0
1	2022	2024	2	2	1	1
1	2024	2025	1	3	0	1
2	2019	NA	2	1	1	1
3	2023	2024	3	1	1	1
3	2024	NA	2	2	1	2
4	2023	NA	1	1	0	0

Pour id=(2,3,4), ce compteur permet d'obtenir l'information souhaitée, à savoir n=0 en situation d'attente/séjour/survie et n=1 l'année de l'évènement. Pour id=1 cependant, l'alternance en y=1 et y=(2,3) ne permet pas de récupérer l'année d'occurrence (première fois en 2 ou 3). Cela peut être fait, en faisant un compteur sur le compteur précédent:

```
df = df %>% group_by(id) %>% mutate(nn = cumsum(n))

kable(df)
```

id	deb	fin	y	nseq	e	n	nn
1	2018	2021	1	1	0	0	0
1	2022	2024	2	2	1	1	1
1	2024	2025	1	3	0	1	2
2	2019	NA	2	1	1	1	1
3	2023	2024	3	1	1	1	1
3	2024	NA	2	2	1	2	3
4	2023	NA	1	1	0	0	0

## Récupération des information censurée

Pour récupérer l'information sur les observations qui seront censurée, on peut faire un total sur la variable *n* ou *e*: si *n*=0, l'individu n'aura pas connu l'évènement.

```
df = df %>% group_by(id) %>% mutate(N = sum(n))
kable(df)
```

id	deb	fin	y	nseq	e	n	nn	N
1	2018	2021	1	1	0	0	0	2
1	2022	2024	2	2	1	1	1	2
1	2024	2025	1	3	0	1	2	2
2	2019	NA	2	1	1	1	1	1
3	2023	2024	3	1	1	1	1	3
3	2024	NA	2	2	1	2	3	3
4	2023	NA	1	1	0	0	0	0

Pour *id*=4, *N* est bien égal à 0.

## Récupération du début de l'exposition

Le début de l'exposition étant ici l'année de début de la première séquence. On peut facilement récupérer cette sur toute les lignes en la repérant (ici en générant une nouvelle variable avec la fonction *ifelse*), et en sommant sa valeur sur les autres lignes (=0).

```
df$ debexp = ifelse(df$nseq==1, df$deb, 0) ①
df = df %>% group_by(id) %>% mutate(debexp = sum(debexp)) ②
kable(df)
```

① La variable *debexp* est égale à *deb* si *nseq*=1, 0 sinon.

② On somme cette valeur sur chaque individu pour l'ajouter aux séquences suivantes.



id	deb	fin	y	nseq	e	n	nn	N	debexp
1	2018	2021	1	1	0	0	0	2	2018
1	2022	2024	2	2	1	1	1	2	2018
1	2024	2025	1	3	0	1	2	2	2018
2	2019	NA	2	1	1	1	1	1	2019
3	2023	2024	3	1	1	1	1	3	2023
3	2024	NA	2	2	1	2	3	3	2023
4	2023	NA	1	1	0	0	0	0	2023

### Mise en forme finale de la base

On peut maintenant conserver les lignes qui nous intéressent à savoir celle où  $nn=1$  (évènement) ou  $N=0$  (censure).

```
df = filter(df, nn==1 | N==0)

kable(df)
```

id	deb	fin	y	nseq	e	n	nn	N	debexp
1	2022	2024	2	2	1	1	1	2	2018
2	2019	NA	2	1	1	1	1	1	2019
3	2023	2024	3	1	1	1	1	3	2023
4	2023	NA	1	1	0	0	0	0	2023

On dispose déjà de la variable d'évènement/censure ( $e$  ou  $n = (0,1)$ ), on finit donc par la variable de durée.

```
df$fin[is.na(df$fin)] = 2025

df$dur = ifelse(df$e==1, df$deb - df$debexp + 1, df$fin - df$debexp + 1)

df = select(df, c(id,e,dur))

kable(df)
```

id	e	dur
1	1	5
2	1	1
3	1	1
4	0	3

Ces informations sont suffisantes pour estimer une fonction de séjour et on peut ajouter, si elles ne sont pas présentes, des covariables fixes issues du fichier des caractéristiques générales. Pour l'ajout

de covariables dynamiques, leur ajout n'est pas forcément difficile pour une analyse en durée discrète<sup>6</sup>. Pour les analyses type Cox, selon la nature de la variable dynamique, l'opération (quel que soit le logiciel utilisé) risque d'être plus ou moins compliquée.

## 11.5 Durée de séjour dans la séquence d'intérêt et variables d'analyse

En première ou deuxième analyse, on peut également voir s'intéresser à la durée de séjour dans l'état précédent. Par exemple, si l'analyse précédent consistait à regarder la durée de séjour dans le premier emploi, on pourrait regarder ensuite la durée jusqu'à sa reprise.

Cela va un peu (voir plus) se compliquer. On va repartir de la base de départ précédente en ajoutant une observation.

```
df = data.frame(id = c( 1, 1, 1, 2, 3, 3, 4, 5, 5, 5 , 5),
  deb = c(2018, 2022, 2024, 2019, 2023, 2024, 2023, 2019, 2021, 2023, 2024),
  fin = c(2021, 2024, 2025, NA, 2024, NA, NA, 2021, 2023, 2024, NA),
  y = c(1, 2, 1, 2, 3, 2, 1, 1, 2, 1, 3),
  nseq = c(1, 2, 3, 1, 1, 2, 1, 1, 2, 3, 4)
)

kable(df)
```

id	deb	fin	y	nseq
1	2018	2021	1	1
1	2022	2024	2	2
1	2024	2025	1	3
2	2019	NA	2	1
3	2023	2024	3	1
3	2024	NA	2	2
4	2023	NA	1	1
5	2019	2021	1	1
5	2021	2023	2	2
5	2023	2024	1	3
5	2024	NA	3	4

### Filtrage des observations hors champs

On peut déjà supprimer les observations hors champs, à savoir ici id=4 qui n'a pas connu l'évènement dont on analyse la durée.

<sup>6</sup>En conservant l'information sur les années, on transformera la base en format individu-période et on procédera à une fusion des informations

```
df$e23 = ifelse(df$y==2 | df$y==3,1,0) ①

df = df %>% group_by(id) %>% mutate(n23 = cumsum(e23))
df = filter(df, n23!=0) ②

kable(df)
```

- ① Nom de la variable *e23* pour repérer la présence de l'évènement dont on analyse la durée.  
 ② Ce compteur est suffisant car l'observation n'a qu'une ligne.

id	deb	fin	y	nseq	e23	n23
1	2022	2024	2	2	1	1
1	2024	2025	1	3	0	1
2	2019	NA	2	1	1	1
3	2023	2024	3	1	1	1
3	2024	NA	2	2	1	2
5	2021	2023	2	2	1	1
5	2023	2024	1	3	0	1
5	2024	NA	3	4	1	2

## Récupération de l'évènement analysé

Ici l'évènement sera un *retour* dans l'état  $y=1$ . Il y a de nouveau une possibilité de censure à droite si une observation reste dans l'état 2 ou 3 jusqu'au moment de l'enquête.

Il peut être utile d'utiliser des variables décalées pour repérer les changements d'état d'une séquence à une autre. Ces décalages sont appelées *lead* ou *lag*:

- **lead**:  $x_t = x_{t+1}$
- **lag**:  $x_t = x_{t-1}$

On va utiliser ici des **lead** et donc pouvoir repérer les changements d'état d'une séquence à une autre. Comme on s'intéresse au retour à l'état 1:

```
df$e = ifelse(df$y==1,1,0) ①

df = df %>% group_by(id) %>% mutate(diff_e = e - lead(e)) ②

kable(df)
```

- ① *e* est une indicatrice qui repère l'état 1  
 ② On fait *redescendre* la valeur de *e* sur la séquence précédente, et on calcule la différence.

id	deb	fin	y	nseq	e23	n23	e	diff_e
1	2022	2024	2	2	1	1	0	-1
1	2024	2025	1	3	0	1	1	NA

id	deb	fin	y	nseq	e23	n23	e	diff_e
2	2019	NA	2	1	1	1	0	NA
3	2023	2024	3	1	1	1	0	0
3	2024	NA	2	2	1	2	0	NA
5	2021	2023	2	2	1	1	0	-1
5	2023	2024	1	3	0	1	1	1
5	2024	NA	3	4	1	2	0	NA

Pour chaque dernière séquence la valeur du lag est une valeur manquante. On repère l'évènement avec une valeur de -1 (transition de 0 à 1). On ne peut pas encore filtrer les informations car il va falloir récupérer la fin de la séquence, mais on peut déjà construire l'information.

```
df$e = ifelse(df$diff_e== -1,1,0)
df$e[is.na(df$e)] = 0
df = df %>% group_by(id) %>% mutate(e = sum(e))

kable(df)
```

id	deb	fin	y	nseq	e23	n23	e	diff_e
1	2022	2024	2	2	1	1	1	-1
1	2024	2025	1	3	0	1	1	NA
2	2019	NA	2	1	1	1	0	NA
3	2023	2024	3	1	1	1	0	0
3	2024	NA	2	2	1	2	0	NA
5	2021	2023	2	2	1	1	1	-1
5	2023	2024	1	3	0	1	1	1
5	2024	NA	3	4	1	2	1	NA

## Récupération de l'année final avec succession d'états de même type

La difficulté ici est apportée seulement par id=3. Jusqu'à 2025, on a successivement l'état 2 puis 3. Il va donc falloir récupérer cette dernière année de succession de 2 et 3, jusqu'à la censure ou jusqu'à un retour dans l'état 1. S'il n'y avait pas ce genre de situation, l'utilisation de la variable *diff\_e* aurait été suffisante pour récupérer l'année de fin lorsqu'on a plusieurs séquences (situations pour id=1,5).

On va de nouveau utiliser un lead, mais sur la variable *e23*.

```
df = select(df, -c(nseq, diff_e)) ①

df = df %>% group_by(id) %>% mutate(lead_e23 = lead(e23, n = 1, default = NA)) ②

df$idem = ifelse(df$e23 == df$lead_e23, 1, 0) ③
df$idem[is.na(df$idem)]=0
df = df %>% group_by(id) %>% mutate(idem = sum(idem)) ④
```

```
kable(df)
```

- ① On supprime les colonnes non utilisées pour gagner ici de la lisibilité
- ② *lead* sur la variable *e23*.
- ③ La variable *idem* permet de repérer une suite d'état 2 et 3. On ne passe pas ici par une variable de différence (le faire par prudence si on le souhaite).
- ④ Ici le total est égal à 1. Si on avait eu une séquence supplémentaire de 3, il serait égal à 2. L'important ici est de repérer la situation, soit 0 ou supérieur à 0.

id	deb	fin	y	e23	n23	e	lead_e23	idem
1	2022	2024	2	1	1	1	0	0
1	2024	2025	1	0	1	1	NA	0
2	2019	NA	2	1	1	0	NA	0
3	2023	2024	3	1	1	0	1	1
3	2024	NA	2	1	2	0	NA	1
5	2021	2023	2	1	1	1	0	0
5	2023	2024	1	0	1	1	1	0
5	2024	NA	3	1	2	1	NA	0

On doit maintenant récupérer la dernière année de fin des situations où *idem*>0, et la placer sur la première.

```
df$fin[is.na(df$fin)] = 2025 ①
df$lead_e23[is.na(df$lead_e23)] = -10 ②

df$truefin = ifelse((df$lead_e23 != df$e23) & df$idem>0, df$fin,0) ③

df = df %>% group_by(id) %>% mutate(truefin = sum(truefin)) ④
df$fin = ifelse(df$idem>0, df$truefin, df$fin)

df = select(df, -c(y,e23,lead_e23,idem))

kable(df)
```

- ① On remplace l'année de la censure par sa valeur (important pour *id*=3).
- ② Pour régler un problème de gestion des NA avec *ifelse*. A tester avec *if\_else* ou *case\_when*.
- ③ On récupère la valeur de l'année de fin lorsqu'il y a une succession d'états de même nature pour l'analyse.
- ④ on remplace la valeur dans la variable *fin* en cas de succession seulement.

id	deb	fin	n23	e	truefin
1	2022	2024	1	1	0
1	2024	2025	1	1	0

id	deb	fin	n23	e	truefin
2	2019	2025	1	0	0
3	2023	2025	1	0	2025
3	2024	2025	2	0	2025
5	2021	2023	1	1	0
5	2023	2024	1	1	0
5	2024	2025	2	1	0

On peut [enfin] sélectionner et conserver une seule ligne par individu et générer la variable de durée

```
df= select(df,-truefin)

df = df %>% group_by(id) %>% mutate(nn23 = cumsum(n23))
df = filter(df, n23==nn23)

df$dur= df$fin - df$deb + 1

df = select(df, -c(n23,nn23))

kable(df)
```

id	deb	fin	e	dur
1	2022	2024	1	3
2	2019	2025	0	7
3	2023	2025	0	3
5	2021	2023	1	3

# 12 Risques concurrents

Le problème des événements multiples dans les analyses de survie a été posé dans les années 1970 avec la notion de **risques concurrents** (*competing risks*) : il s'agit d'événements survenant au cours de la période d'observations et qui *empêchent* l'occurrence de l'événement d'intérêt.

## 12.1 Problématique

On étudie un processus dont l'occurrence a plusieurs modalités, *types* ou *causes*:

- 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
- Les types de rupture d'union: séparation-divorce, veuvage).

Rappel: Déjà abordé dans la partie théorie, avec un recueil de données de type prospectif les “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 de Kaplan-Meier n'est plus une probabilité. Une estimation de type KM d'un événement en concurrence avec d'autres impose que ces derniers soient traités comme des censures à droites non informatives. Mais il n'est pas possible de tester cette hypothèse.

## 12.2 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 événement unique. Par exemple les décès si on analyse ses causes.

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

### Cause specific risk

Pour le risque de type  $k$ , le risque *cause-spécifique* 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

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 sur cette information. Une sortie concurrente a donc été simulée sans plus de précision (variable *compet*), que l'on considèrera non strictement indépendante à la cause d'intérêt. Ce risque entre donc en concurrence avec la cause du décès directement liée à la malformation cardiaque, que la personne ait été transplantée ou non.

<IPython.core.display.HTML object>

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

Variable *compet*:

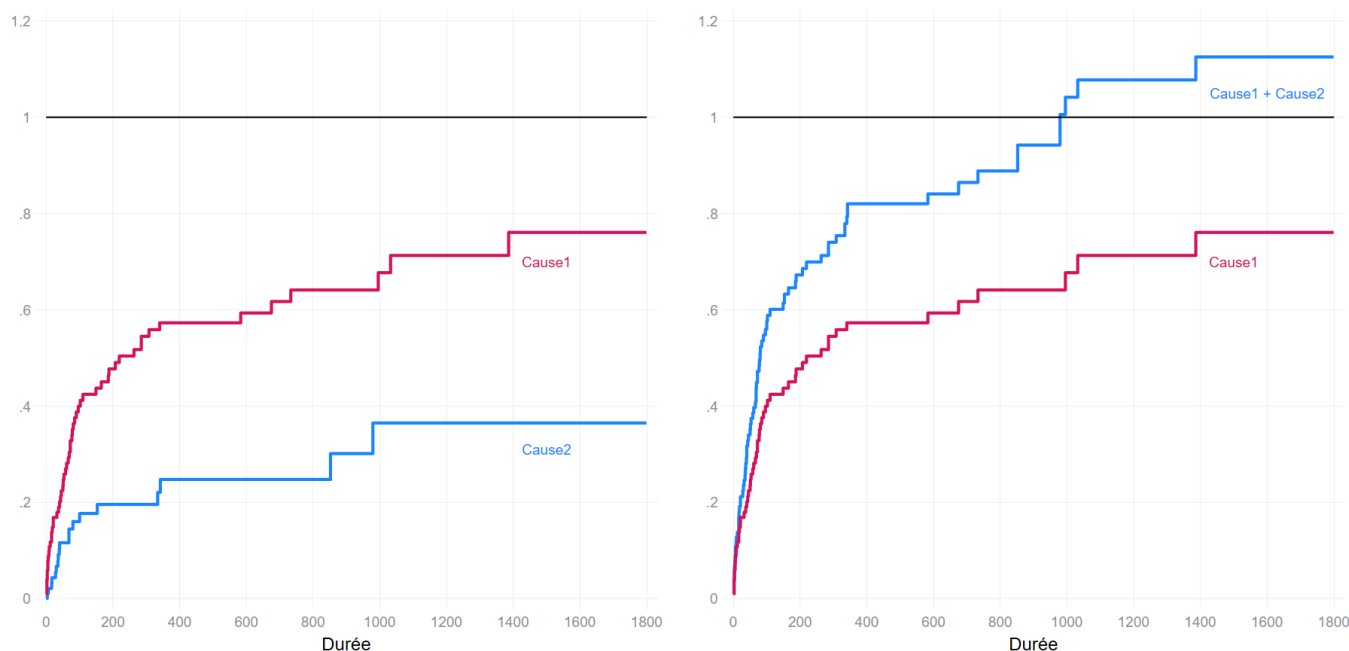
cause 1 => décès directement provoquer par la malformation: *compet*=1 cause 2 => autre cause  
*compet*=0 => censure à droite

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 ( $n = 18 + 29 = 48$ ), puis en sommant les deux estimateurs, la fonction de répartition excède 100% au bout de 1000 jours environs. La proportion de survivant.e.s est donc négative.



Figure 12.1: Fonction de répartition avec une cause concurrente traitée comme une censure à droite



## 12.3 Estimations en présence de risques concurrents (CIF)

### 12.3.1 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 de risques concurrents dans une logique dite *cause specific*.

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

- Actuellement, l'estimateur le plus utilisé est la fonction dite d'**incidence cumulée - CIF-** de Kalbfleisch-Prentice et 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* (R, SAS) ou *test de Pepe-Mori* (Stata).

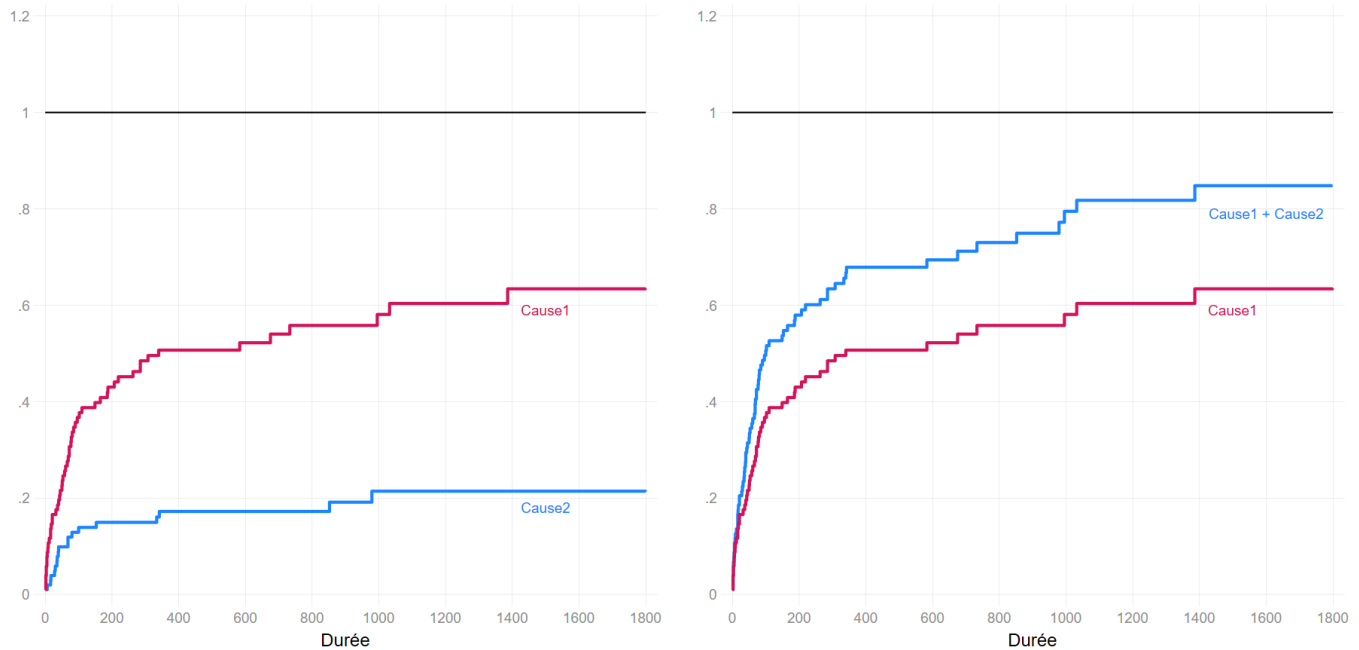
### CIF (Cumulative Incidence Function)

- Si  $h_k(t_i)$  est le risque *cause-spécific* 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, l'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 pour la cause  $k$  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  $S(t_i - 1)$ .
- 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  $k$ .

Figure 12.2: Risques concurrent: estimation de la CIF



```

failure: compet == 1
competing failures: compet == 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
1032	0.6036	0.0559	0.4851	0.7031
1386	0.6340	0.0583	0.5083	0.7357

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

En présence du risque concurrent, et traité comme tel, la moitié des personnes sont décédées suite à 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 du quotient d'intensité dite « cause-specific ».

### 💡 R-Stata-Sas-Python

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.

Pour l'estimation des CIF (risque de sous répartition):

- **R:** la librairie **cmprsk** permet d'estimer simplement les incidences cumulées avec la fonction **cuminc**.
- **Sas:** maintenant directement estimable avec **proc lifetest**. Il suffit d'indiquer le ou les risques d'intérêt dans l'instruction indiquant la variable de durée et de censure avec l'option **failcode=valeur**.
- **Stata:** Estimation avec la commande externe **stcomp**. La commande génère des variables qui demande des manipulations supplémentaires pour afficher les résultats sous forme de tableau par exemple. On peut utiliser et préférer la commande externe **stcomlist**.
- **Python:** le wrapper de R (**cmprsk**) ne fonctionne plus à ce jour à défaut de mise à jour [2022].

## 12.3.2 Compararaison des CIF

- **Test d’homogénéité de Gray**: est basé sur une autre mesure du risque en évènement concurrent. Sur le principe, identique à la philosophie des test du logrank. 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 et R. 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. Je n’ai pas le recul nécessaire sur cette alternative, qui n’est implémenté que dans Stata.

Table 12.1: Test de Gray pour la variable surgery

Risques	Chi2	P>Chi2
Cause1	5.783	0.0161
Cause2	0.129	0.7191

Table 12.2: Test de Pepe-Mori pour la variable surgery

Risques	Chi2	P>Chi2
Cause1	6.203	0.0127
Cause2	1.880	0.7038

### 💡 R-Stata-Sas-Python

- **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`: non testée) :
- **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 `stpepemori`.
- **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.

## 12.4 Modèles

### 12.4.1 Modèles Semi paramétriques

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.

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

Modèle de Cox «standard» pour chaque évènement, 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. Il ne sera donc pas exécuté pour l'application

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>

#### R-Stata-Sas-Python

- **R**: on utilise la fonction **crr** du package **cmprsk**.
- **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**.
- **Python** : ne pas essayer d'utiliser la librairie **cmprsk** qui n'est pas mis à jour et ne fonctionne donc plus.

### 12.4.2 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 traités comme des censures à droite.

Table 12.3: Modèle logistique multinomial avec risques concurrent

(a) Cause 1				(b) Cause 2			
Cause 1	RRR	p> z	95% IC	Cause 2	RRR	p> z	95% IC
<i>t</i>	0.816	0.000	0.752 - 0.885	<i>t</i>	0.817	0.003	0.713 - 0.935
<i>t</i> <sup>2</sup>	1.003	0.000	1.001 - 1.005	<i>t</i> <sup>2</sup>	1.003	0.052	1.000 - 1.006
<i>year</i>	0.879	0.116	0.749 - 1.032	<i>year</i>	0.816	0.141	0.622 - 1.070
<i>age</i>	1.045	0.012	1.010 - 1.081	<i>age</i>	1.011	0.654	0.964 - 1.061
<i>surgery</i>	0.318	0.033	0.110 - 0.913	<i>surgery</i>	0.541	0.431	0.117 - 2.496
<i>constante</i>	0.231	0.000	0.148 - 0.360	<i>constante</i>	0.076	0.000	0.037 - 0.157

- 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 événements qui sont supposés non indépendants. Néanmoins la dépendance entre risques concurrents n'est pas non plus stricte et cette hypothèse d'IIA, seulement testable par le bon sens, est souvent illustrée par l'exemple des couleurs des bus dans le choix du mode de transport, ou les couleurs de chaussure dans les études marketing. Soit est une situation relativement limite.
- En terme de lecture, les estimateurs du modèle logistique multinomial peuvent directement s'interpréter comme des rapports de risque (ou relative risk ratio).
- 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, on peut utiliser une méthode de standardisation, de type **AME** (*Average Marginal Effect*).

Pour l'application, nous avons pris le mois (30 jours) comme métrique temporelle. On rappelle que les valeurs des estimateurs sont fictives en raison de la simulation des événements pour le risque concurrent (cause2)

Notes:

- On a utilisé le terme RRR - Relative Risk Ratio - pour la colonne rapportant les estimations. Dans un cadre de risque concurrent il est un peu difficile d'utiliser formellement la notion de *hazard rate* tel qu'il a été défini plus haut, enfin les modèles multinomiaux ne reportent pas formellement des Odds Ratios dont l'utilisation devrait être réservée exclusivement à une alternative binaire.
- les variables *year* et *age* ont été centrées sur leur valeur moyenne pour donner aux constantes des valeurs acceptables.
- Pour faciliter la lecture on peut utiliser une méthode de standardisation de type AME (*Average Marginal Effect*).

# 13 Modèles paramétriques

**Objectifs:** présenter, assez rapidement, la logique des modèles de type **AFT** (**Accelerated Failure Time**), principalement celui de *Weibull*.

## 13.1 Principes

- Dans les modèles paramétriques usuels<sup>1</sup>, la durée de survie est distribuée selon une loi dont la densité  $f(t)$  pleinement paramétrée.
- Pour utiliser l'approche paramétrique, il faut avoir de bonnes raisons de penser que les durées de survie sont distribuées selon une certaine loi connue plutôt qu'une autre.
- La majorité des distributions reposent sur une hypothèse dite AFT (la proportionnalité des risques. Certaines modèles peuvent reposer sur les deux comme le modèle de *Weibull*. Enfin, les modèles *log-logistique* ou *log-normal* n'ont qu'une paramétrisation de type *AFT*. Depuis une vingtaine, un modèle dit *flexible* (Parmar-Royston) ne paramétrise pas la distributions des évènements à partir d'une loi mais à partir d'un ajustement reposant sur la méthode de lissage par spline cubique<sup>2</sup>.

## 13.2 Hypothèse AFT: Accelerated Failure Time

L'hypothèse **AFT** signifie que l'effet des covariables est multiplicatif par rapport à la durée de survie/séjour. 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 donc plus la même métrique temporelle. Cela renvoie à des interprétations de type *dilation/contraction* du temps, par analogie à la théorie de la relativité, mais ici avec une seule dimension.

Exemple simple: la durée de vie d'un être humain et d'un chien.

On dit qu'une année de vie d'un être humain est équivalent 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.

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

---

<sup>1</sup>Plus anciens que le modèle de Cox

<sup>2</sup>permet un ajustement relativement proche des modèles à durée groupée mais avec des durées strictement continues



Remarque: si un modèle s'estime AFT s'estime également sous hypothèse PH, comme celui de 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. Cela s'avère donc pratique après une analyse non paramétrique de type Kaplan-Meier par exemple.
- Inconvénient: très difficile d'introduire de variables dynamiques<sup>3</sup>.

*Humain versus 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'écoulerait donc plus vite pour le chien que pour l'être humain du point de vue d'un référentiel extérieur. 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 la durée de survie est plus courte, alors le risque est plus élevé.

### 13.3 Principe de construction des modèles AFT

Le raisonnement mathématique est ici plus complexe que pour les modèles de Cox ou à durée discrète. 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 temps continu (pas d'évènement simultané):

$$\log(t_i) = \alpha_0 + \alpha_1 X_i + bu_i$$

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

On peut écrire:  $f(u_i) = f(\frac{\log(t_i) - \alpha_0 - \alpha_1 X_i}{b})$ .

Remarque: pour une distribution normale/gaussienne, le paramètre de position est l'espérance et le paramètre d'échelle l'écart-type.

### 13.4 Quelques modèles paramétriques usuels

Modèle exponentiel et de Weibull

Weibull

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

---

<sup>3</sup>Jamis testé de mon côté

Table 13.1: Modèle de Weibull

(a) Accelerated Failure Time (AFT)				(b) Proportional hazard (PH)			
Variables	Time Ratio	$p >  z $	95% IC	Variables	HR	$p >  z $	95% IC
<i>year</i>	1.176	0.184	0.926 - 1.493	<i>year</i>	0.914	0.175	0.802 - 1.041
<i>age</i>	0.940	0.013	0.896 - 0.987	<i>age</i>	1.035	0.014	1.007 - 1.063
<i>surgery</i>	7.173	0.011	1.557 - 33.048	<i>surgery</i>	0.334	0.012	0.143 - 0.783
$\rho$	0.556	-	0.464 - 0.667	$\rho$	0.556	-	0.464 - 0.667

### Exponentiel

- Processus sans mémoire, utilisé pour étudier par exemple la durée de vie 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 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).

### 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 (PH seulement), Gamma et Gamma généralisé.....

**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 regarder s'il ajuste bien les données si la transformation  $\log(-\log(S(t_i)))$  est linéaire par rapport à  $\log(t_i)$ .

### Application

#### Comparaison des AIC (sans covariable)

- Weibull: 400.1
- Exponentiel: 461.0
- Gompertz: 409.6
- Log-logistique: 391.8

## 13.5 Exemple avec le modèle de Weibull

Note: la constante n'est pas reporté.  $\rho$  indique la valeur estimé d'un paramètre de *forme*. Son signe indique sur le risque est décroissant ou croissant (1 si risque constant), et permet de passer de la paramétrisation AFT à la paramétrisation PH (et inversement).

- **AFT**: Un jour de survie d'une personne qui n'a pas été opérée d'un pontage correspond environ à 7 jours de survie d'une personne opérée. Cette remise à l'échelle de la métrique temporelle entre les deux groupes exprime bien le gain en durée de survie pour les personnes opérées, soit des risques journaliers de décès plus faibles (et plus faibles à valeurs constantes, proportionnalité oblige).
- **PH**: Lecture en rapport de risque ou *hazard rate* (idem Cox). Si on avait reporté les coefficients (échelle log)  $b_{ph} = -\rho \times b_{aft}$ . Ici  $-0.556 \times (1.97) = -1.096$ . Et  $e^{-1.096} = 0.334$

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 diffèrent.

## 13.6 Le modèle de Parmar-Royston

- Le bon ajustement par une loi de distribution prédéfinie peut s'avérer contraignante. Le modèle de Cox avait justement pour objectif de se défaire de cette contrainte, la plupart des distributions utilisées étant monotone ou unimodale (log-logistique ou log-normal).
- Le principe des splines peut-être rapproché de celui qui a été utilisé plus haut dans le modèle logistique à durée discrète avec l'introduction des polynômes [  $f(t) = (a_1 \times t) + (a_2 \times t^2) + (a_3 \times t^3) + \dots + (a_k \times t^k)$ .
  - Cette méthode brute d'ajustement consiste finalement à introduire une interaction ou plusieurs interactions entre la variable de durée elle-même.
  - L'ajustement par des polynômes classiques est très sensible aux outliers (overfitting) au delà de  $k > 2$  <sup>4</sup>.
- Du côté des **splines cubiques** la méthode d'ajustement et de lissage est de meilleure qualité et permet de mieux contenir les problèmes d'overfitting.
  - les splines cubiques sont donc basées sur des polynômes d'ordre 3 (d'où cubique) avec une estimation par morceau (intervalles). les morceaux sont définis manuellement ou par un nombre de degrés de liberté obtenu à partir des quantiles du logarithme de la fonction de survie après avoir exclu les observation censurées.
    - \* Deux degrés de liberté (1 noeud) avec un intervalle allant jusqu'au log de la moitié des survivants et un second à partir de cette seconde moitié.
    - \* Sur le même principe trois degrés de liberté (2 noeuds) coupe la durée en 3 intervalles sur ses terciles.
    - \* En pratique, il est préférable de donner à l'application le nombre de degré de liberté plutôt que d'indiquer manuellement la position des noeuds.

---

<sup>4</sup>lors la formation il suffit de la tester avec la base des TP pour  $k=3$  et calculer la probabilité conditionnelle pour s'en convaincre

- \* Il convient également de ne pas être trop gourmand sur le nombre de noeuds, un ou deux étant souvent suffisant (donc 2 ou 3 degrés de liberté).
- \* On peut choisir le nombre de degrés de liberté en estimant des modèles sans covariable et comparer les AIC (vraisemblance pénalisée).
- Contrairement aux autres modèles, et sans rentrer dans les détails, le modèle de Parmar-Royston part de la fonction de risque cumulée et non des taux de risque/hasard. Les risk ratios sont obtenus en utilisant les relations entre les différentes grandeurs (voir section *théorie*).

## Exemple

Avec 2 degrés de liberté (un noeud):

Table 13.2: Modèle de Parmar-Royston

Variables	$\hat{e}(b)$	$p >  z $	95% IC
<i>year</i>	0.885	0.067	0.777 - 1.008
<i>age</i>	1.030	0.026	1.004 - 1.058
<i>surgery</i>	0.373	0.025	0.159 - 0.876
<i>spline1</i>	3.157	0.000	2.503 - 3.981
<i>spline2</i>	1.289	0.002	1.099 - 1.511
<i>constante</i>	0.510	0.000	0.386 - 0.674

A savoir:

- Avec un degré de liberté, le modèle de Parmar-Royston estime un modèle de Weibull sous paramétrisation PH.
- Les paramètres pour les splines ne sont pas interprétables directement. Ils servent à calculer la baseline du risque via l'équation du polynôme (non reporté car expression bien corsée).
- De nouveau il s'agit d'un modèle à risque proportionnel.

# 14 Annexes

## 14.1 Tests Grambsch-Therneau OLS sur les résidus de Schoenfeld

### ! Important

Attention il ne s'agit pas du test actuellement implémenté dans la nouvelle version de `survival` (v3) qui, malheureusement, lui a substitué la version dite *exacte* (moindres carrés généralisés). Le programme de la fonction du test OLS est néanmoins facilement récupérable et exécutable. [lien](#). Je continue de préconiser l'utilisation de cette version OLS du test, reproductible avec les autres applications statistiques (Stata,Sas,Python).

- Le test dit “simplifié”, qui n'apparaît pas dans le texte original de P.Gramsch et T.Thernau [lien](#), répond à un souci d'instabilité des variances des résidus de Schoenfeld en fin de durée d'observation lorsque peu d'observation restent soumises au risque. Cet argument est soulevé dans leur ouvrage de 2022 [lien](#) avant d'en présenter sa version.
- Il est simplifié car on applique à tous les résidus bruts la variance du paramètre ( $b$ ) estimés par le modèle de Cox.
- Le test devient alors un simple test de corrélation entre les résidus et une fonction de la durée (centrée). Dans l'esprit, il peut être également approché par une régression linéaire par les moindres carrés ordinaires entre les résidus et une fonction de la durée (voir page 134 de l'ouvrage de Grambsch et Therneau).

Soit les données suivantes, avec  $t$  la variable de durées,  $Y$  la variable de censure et  $X$  la seule et unique covariable.

- Pas d'évènement simultané (donc pas de correction de la vraisemblance)
- Covariable de type indicatrice

$t_i$	$Y_i$	$X_i$
1	1	1
2	0	0
3	0	0
4	1	1
5	1	1
6	1	0
7	0	1

```
test = data.frame(time= c(1,2,3,4,5,6,7),
                  Y=c(1,0,0,1,1,1,0),
                  X=      c(1,0,0,1,1,0,1))
```

Estimation du modèle de Cox:

```
library(survival)
fit = coxph(formula = Surv(time, Y) ~ X, data=test)
fit
```

Call:

```
coxph(formula = Surv(time, Y) ~ X, data = test)
```

```
      coef exp(coef) se(coef)      z      p
X 0.6217    1.8622    1.1723 0.53 0.596
```

```
Likelihood ratio test=0.31  on 1 df, p=0.5797
n= 7, number of events= 4
```

Calcul des résidus brut (si et seulement si  $Y = 1$ ) dans le cas d'une seule covariable avec  $b$  égal à **0.62**:

$$rs_i = X_i - \sum_{j \in R_i} X_i \frac{e^{0.62 \times X}}{\sum_{j \in R_i} e^{0.62 \times X}} = X_i - E(X_{j \in R_i})$$

Il y a ici 4 résidus à calculer, pour  $t = (1, 4, 5, 6)$

**Résidus pour  $t = 1$**

- $a_1 = \sum_{j \in R_i} e^{0.62 \times X} = e^{0.62} + 1 + 1 + e^{0.62} + 1 + e^{0.62} = 10.43$
- $b_1 = \sum_{j \in R_i} X_i \frac{e^{0.62 \times X}}{\sum_{j \in R_i} e^{0.62 \times X}} = 4 \times \frac{e^{0.62}}{10.43} = 0.71$
- $r_1 = 1 - 0.71 = 0.29$

**Résidus pour  $t = 4$**

- $a_4 = e^{0.62} + e^{0.62} + 1 + e^{0.62} = 6.58$
- $b_4 = 4 \times \frac{e^{0.62}}{6.58} = 0.84$
- $r_4 = 1 - 0.84 = 0.15$

**Résidus pour  $t = 5$**

- $a_5 = e^{0.62} + e^{0.62} + 1 = 4.71$
- $b_5 = 2 \times \frac{e^{0.62}}{4.71} = 0.78$
- $r_5 = 1 - 0.78 = 0.21$

**Résidus pour  $t = 6$**

- $a_6 = e^{0.62} + 1 = 2.86$

- $b_6 = \frac{e^{0.62}}{2.86} = 0.65$
- $r_6 = 0 - 0.65 = -0.65$

Les résidus “standardisés”, ou plutôt *scaled residuals* (je cale sur une traduction correcte en français) sont égaux à:

$$sr_i = b + nd \times Var(b) \times r_i$$

Avec  $nd = \sum Y_i$

- $\sum Y_i = 4$
- $Var(b) = (1.1723)^2 = 1.37$
- $sr_1 = 0.62 + 4 \times 1.37 \times 0.29 = 2.20$
- $sr_4 = 0.62 + 4 \times 1.37 \times 0.15 = 1.47$
- $sr_5 = 0.62 + 4 \times 1.37 \times 0.21 = 1.78$
- $sr_6 = 0.62 + 4 \times 1.37 \times (-0.65) = -2.95$

Avec  $g(t_i)$  une fonction de la durée ( $g(t_i) = t_i$ ,  $g(t_i) = 1 - KM(t_i)...$ ) et  $\overline{g(t)}$  sa valeur moyenne, la statistique du test score simplifié pour une covariable est égale à :

$$\frac{[\sum_i (g(t_i) - \overline{g(t)}) \times sr_i]^2}{nd \times Var(b) \times (\sum_i (g(t_i) - \overline{g(t)})^2)}$$

Et suis un  $\chi^2$  à 1 degré de liberté.

Avec  $\overline{g(t_i)} = t_i$ , le calcul de la statistique de test est:

- $\overline{g(t_i)} = \frac{28}{7} = 4$
- $\frac{[(1-4) \times 2.20] + [(4-4) \times 1.47 + (5-4) \times 1.78 + (6-4) \times (-2.95)]^2}{4 \times 1.37 \times [(1-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2]} = \frac{114.9}{76.72} = 1.49$

```
#source("D:/D/Marc/SMS/FORMATIONS/analyse_duree/cox.zphold/cox.zphold.R")
```

```
source("https://raw.githubusercontent.com/mthevenin/analyse_duree/master/cox.zphold/cox.zphold.R")
```

```
cox.zphold(fit, transform="identity")
```

Warning in is.R(): 'is.R' est obsolète.

Voir help("Deprecated") et help("base-deprecated").

```
      rho chisq      p
X -0.688  1.49 0.222
```

## 14.2 Fragilité et immunité

Seulement quelques remarques, le traitement de ces problématiques dépassant largement le contenu de la formation.

### 14.2.1 Fragilité (Frailty)

Pour la *fragilité*, je conseille fortement de lire la dernière section du document de travail de **Simon Quantin** (cf bibliographie), *il n'y a pas meilleure présentation du problème que la sienne* [petite maj par rapport à la version précédente: il ne traite que la fragilité individuelle stricto sensu et non la fragilité plus connu sous le terme de shared frailty\* (proche modèle multiniveau). Problématique importante, car une des origines de la non proportionnalité des risques réside dans l'omission de variables. Ici on va être confronté une omission sur des traits non observables ou latents, qui **accélèrent** 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, et rendant l'interprétation des modèles plus compliquée.

On peut distinguer deux types de modèles:

- les modèles à *fragilité partagée*, c'est la situation la plus simple car la logique se rapproche des modèles multiniveaux, des groupes d'individus, identifiables, partagent une même *fragilité*, par exemple géographique.
- les modèles à *fragilité non partagée*, avec des caractéristiques latentes non observable comme les préférences, ou en médecine certains traits génétiques non identifiés.

### 14.2.2 Immunité (Cure fraction)

Le phénomène d'immunité est un cas particulier du précédent, et a été étudié dès le début des années 1950, en questionnant l'exposition au risque d'une partie des observations. On s'interrogeait par exemple sur les risques de rechute et de décès après le traitement d'un premier cancer. Visuellement on peut commencer à se poser des questions sur la présence d'une fraction *immunisée* ou non *susceptible* de connaître l'évènement lorsque la fonction de séjour ne tend pas vers 0 mais présente une longue asymptote (plateau) sur une valeur supérieure à 0:  $\lim_{t \rightarrow \infty} S(t) = a$ .

Les modèles avec une fraction immunisée peuvent être de **type mixte** en associant une probabilité d'être immunisé aux observations censurées à droite à un modèle de durée <sup>1</sup>. Plus dans le vent je crois, on a également des modèles de **type non mixte**, avec il me semble une connotation bayésienne qui semble s'accroître. Il n'y a donc pas de méthode unifiée à ce jour [Si vous voulez vous en convaincre].

---

<sup>1</sup>Le plus classique utilise un algorithme *Expectation Maximisation* utilisé en imputation: on estime une probabilité d'être susceptible de connaître l'évènement aux observations censurées à droite, qui intervient comme facteur de pondération dans le modèle de durée. Cette probabilité et le modèle de durée qui lui est associé est réévalué à chaque boucle de l'algorithme jusqu'à convergence. Le principal problème de cette méthode réside dans l'estimation de la variance, souvent effectué par bootstrap. Cette méthode a l'avantage d'être implémentable en durée discrète, bien qu'à ma connaissance aucun logiciel ne la propose (j'ai une commande Stata encore perfectible sous le coude). On trouve en revanche ce type d'estimation sous R, pour les modèles de Cox ou les modèles paramétriques dans le package **smcure**



On peut également noter, c'est important, que cette problématique affecte les analyses avec des événements dits récurrents. Ici, la stratégie classique qui consiste à introduire dans un modèle un simple effet aléatoire de type fragilité partagée (shared frailty) pour contrôler risque d'être insuffisante. Ici le *groupe* est constitué de chaque séquence de remise dans le risque set. Exemple pour la fécondité: une personne ayant eu un enfant est exposée au risque d'en avoir un autre, l'horloge temporelle étant alors simplement réinitialisée. Et donc, quid des préférences individuelles en terme de fécondité <sup>2</sup>.

Enfin, les modèles à fragilité ou à fraction immunisée repose tous sur une hypothèse très forte. La fragilité ou le degré d'immunité est toujours défini (estimé) en début d'exposition, et il ne varie pas. Cela peut ne pas toujours faire sens, en particulier pour les préférences, pas forcément stables ou fixes dans le temps.

---

<sup>2</sup>en situation de récurrence, toujours penser à *remettre à jour* les conditions initiales, par exemple pour la fécondité l'âge de la mère à la naissance de l'enfant pour les rang supérieur à 1

**partie VI**

**Programmation**

# 15 R

Programme de cette section: [Lien](#)

## 15.1 Packages et fonctions

Analyse	Packages - Fonctions
Non paramétrique	<ul style="list-style-type: none"><li>• discsurv<ul style="list-style-type: none"><li>– lifetable</li><li>– contToDisc</li></ul></li><li>• survival<ul style="list-style-type: none"><li>– survfit</li><li>– survdif</li></ul></li><li>• survRM2<ul style="list-style-type: none"><li>– rmst2</li></ul></li></ul>
Modèles à risques proportionnel	<ul style="list-style-type: none"><li>• survival<ul style="list-style-type: none"><li>– coxph</li><li>– cox.zph (v3) cox.zphold (récupération v2)</li><li>– survsplit</li></ul></li><li>• base et tydir<ul style="list-style-type: none"><li>– uncount</li><li>– glm</li></ul></li></ul>
Modèles paramétriques (ph ou aft)	<ul style="list-style-type: none"><li>• survival<ul style="list-style-type: none"><li>– survreg</li></ul></li><li>• flexsurv<ul style="list-style-type: none"><li>– survreg</li></ul></li></ul>
Risques concurrents	<ul style="list-style-type: none"><li>• cmprsk<ul style="list-style-type: none"><li>– cuminc</li></ul></li><li>• nnet<ul style="list-style-type: none"><li>– multinom</li></ul></li></ul>

Analyse	Packages - Fonctions
Autres (graphiques - mise en forme)	<ul style="list-style-type: none"> <li>• survminer</li> <li>• jtools</li> <li>• gtsummary</li> </ul>

## Installation

Les dernières versions de certains packages peuvent être installées via Github (ex: **survminer**). Pour les récupérer, passer par le package **devtools**.

```
#install.packages("survival")
#install.packages("survminer")
#install.packages("flexsurv")
#install.packages("survRM2")
#install.packages("tidyr")
#install.packages("dplyr")
#install.packages("jtools")
#install.packages("gtools")
#install.packages("cmprsk")
#install.packages("gtsummary")
#install.packages("muhaz")
#install.packages("nnet")
```

```
library(survival)
library(survminer)
library(flexsurv)
library(survRM2)
library(tidyr)
library(dplyr)
library(jtools)
library(gtools)
library(cmprsk)
library(discSurv)
library(gtsummary)
library(muhaz)
library(nnet)
```

## 15.2 Analyse Non paramétrique

Chargement de la base transplantation

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

### 15.2.1 Méthode actuarielle

La fonction disponible du paquet `discsurv`, `lifetable()`, a des fonctionnalités plutôt limitées. Si on peut maintenant définir des intervalles de durée, il n'y a toujours pas d'estimateurs les différents quantiles de la courbe de survie, ce qui limite fortement son utilisation.

La programmation est rendue un peu compliquée pour pas grand chose. Je donne les codes pour information, sans plus de commentaires.

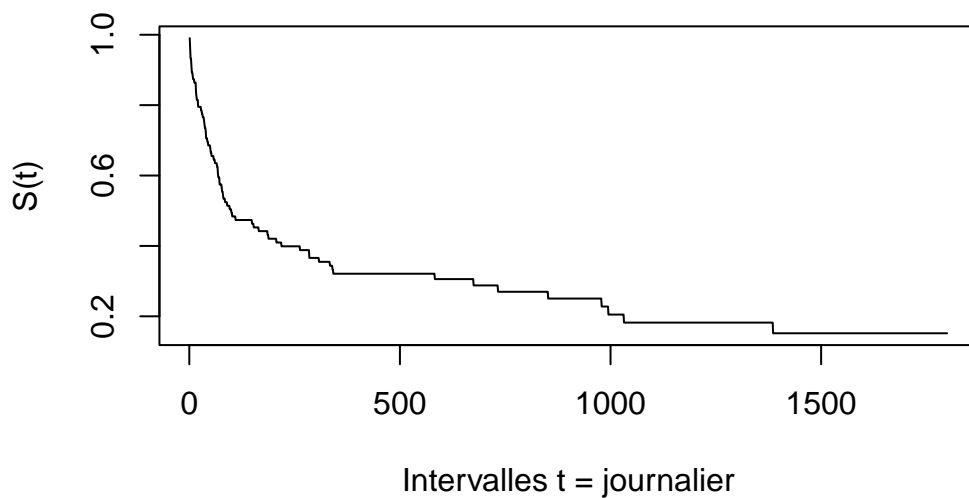
```
trans = as.data.frame(trans)
```

#### Fonction lifeTable

*Intervalle par défaut  $dt = 1$*

```
lt = lifeTable(dataShort=trans, timeColumn="stime", eventColumn = "died")
plot(lt, x = 1:dim(lt$Output)[1], y = lt$Output$S, xlab = "Intervalles t = journalier", ylab=
```

Figure 15.1:  $S(t)$  méthode actuarielle avec `discSurv` (1)



*Intervalle  $dt = 30$*

```
# On définit un vecteur définissant les intervalles (il n'y avait pas plus simple???)
dt <- 1:ceiling(max(trans$time)/30)*30

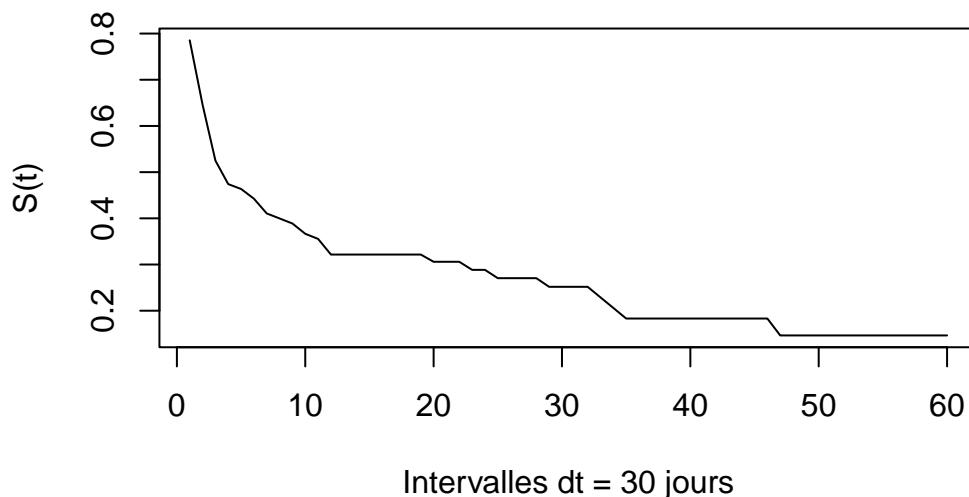
# Base dis avec une nouvelle variable de durée => timeDisc

dis <- contToDisc(dataShort=trans, timeColumn="time", intervalLimits = dt )

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

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

Figure 15.2: Méthode actuarielle avec discSurv (2)



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

### 15.2.2 Méthode Kaplan-Meier

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

#### Estimation des fonctions de survie

Fonction **survfit**

On peut renseigner directement les variables permettant de calculer la durée et non la variable de durée elle-même. Cette méthode est utilisée lorsqu'on introduit une variable dynamique dans un modèle semi-paramétrique de Cox (**coxph**).

---

**Listing 15.1 Syntaxe**

---

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

---

---

**Listing 15.2 Syntaxe**

---

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

---

Sans comparaison de groupes:

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

fit
```

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

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

```
summary(fit)
```

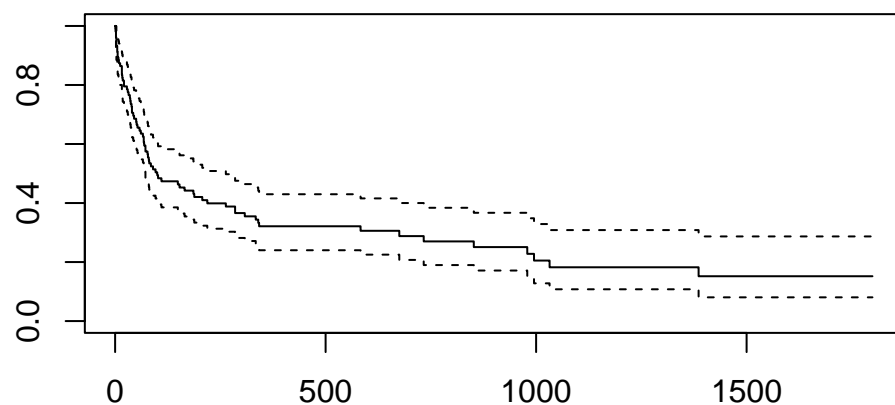
Call: survfit(formula = Surv(stime, died) ~ 1, data = trans)

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

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



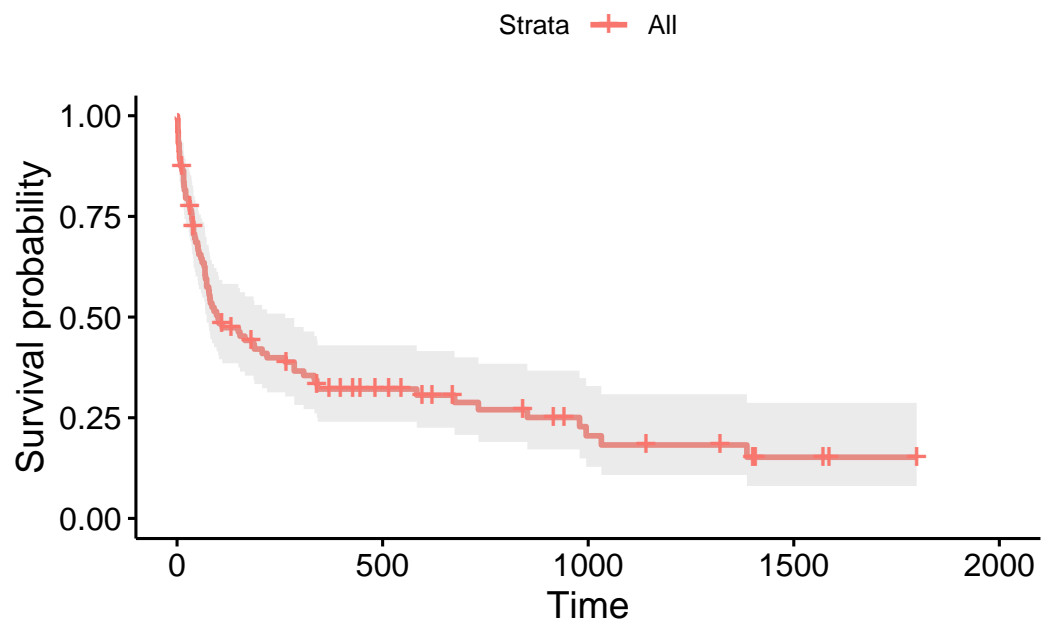
```
plot(fit)
```



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

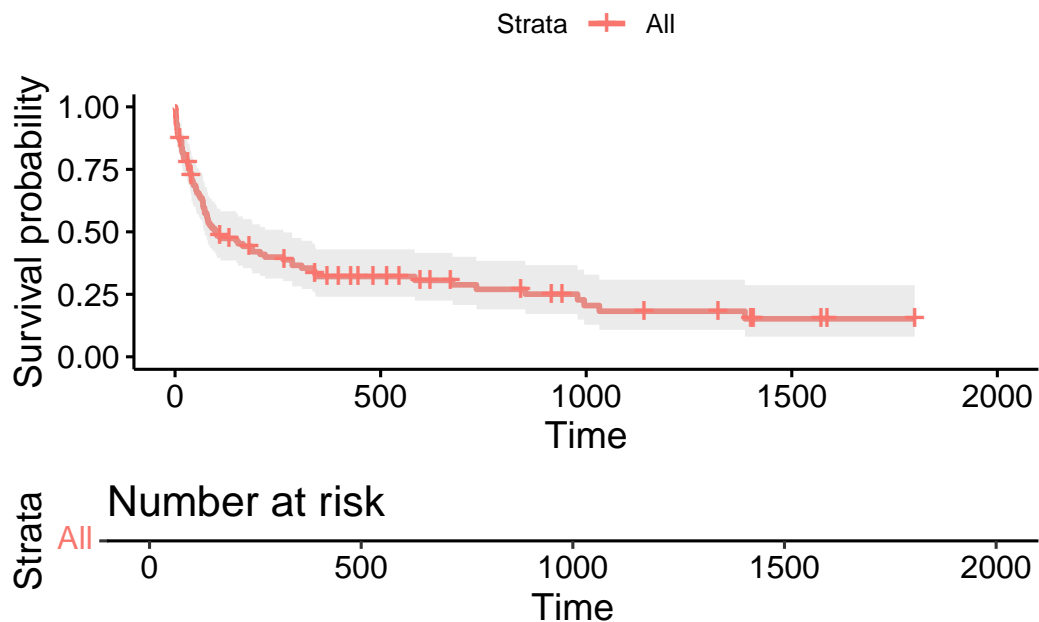
On peut obtenir des graphes de meilleur qualité avec la librairie `survminer`, avec la fonction `ggsurvplot`

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



On peut ajouter la population encore soumise au risque à plusieurs points d'observation avec l'argument `risk.table = TRUE`

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



### 15.2.3 Comparaison des $S(t)$ méthode KM

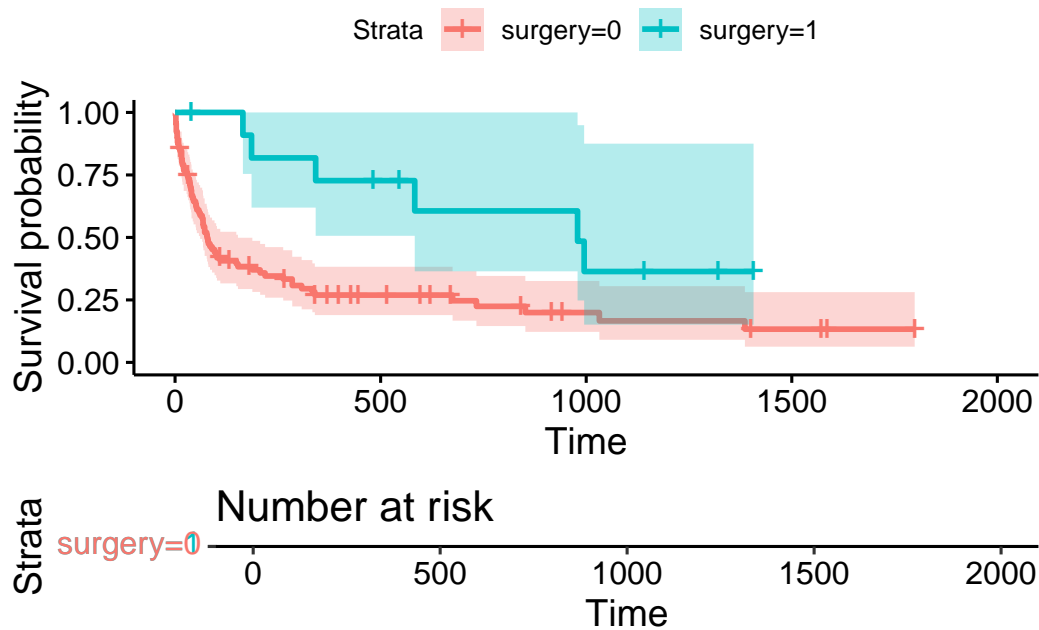
On va comparer les deux fonctions de séjour pour la variable *surgery*, celle pour les personnes non opérées et celle pour les personnes opérées.

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

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

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

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



### Tests du logrank

On utilise la fonction **survdif**, avec comme variante le test des frères Peto ( $\rho=1$ ). La syntaxe est quasiment identique à la fonction **survdif**.

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

Call:

```
survdif(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 veut tester deux à deux les niveaux d'une variable catégorielle à plus de deux modalités, il est fortement conseillé d'utiliser la fonction **pairwise\_survdif** de **survminer** (syntaxe identique que **survdif**).

### Comparaison des RMST

La fonction **rmst2** du package **survRM2** permet de comparer les RMST entre 2 groupes. La strate pour les comparaisons doit être impérativement renommée *arm*. La fonction, issue d'une commande de Stata, n'est 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

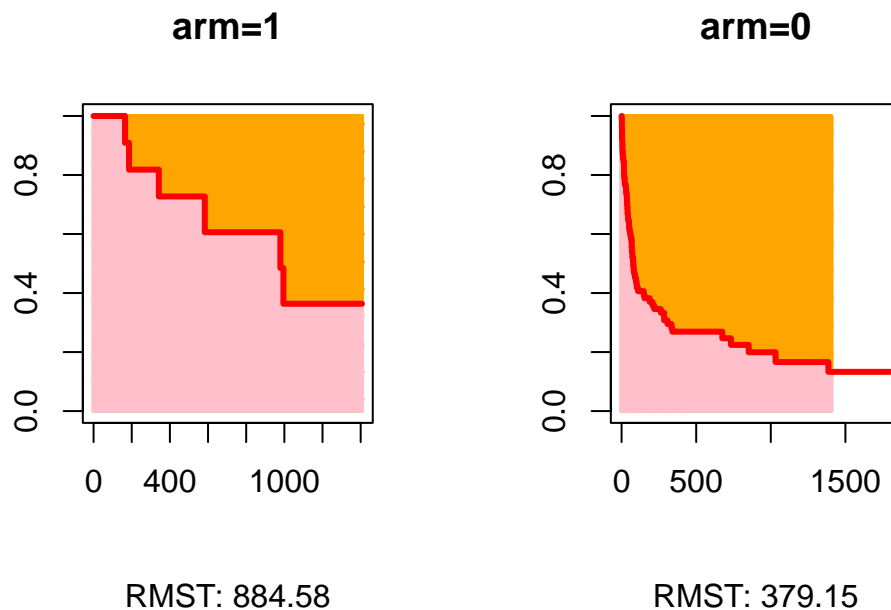
Restricted Mean Time Lost (RMTL) by arm

	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)
```



## 15.3 Modèle de Cox

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

- Estimation du modèle: `coxph`.
- Test de Grambsch-Therneau: `cox.zph` et `cox.oldzph`.
- Introduction d'une variable dynamique: allongement de la base avec `survsplit`.

### 15.3.1 Estimation du modèle

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

Syntaxe:

---

#### Listing 15.3 Syntaxe

---

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

---

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

Call:

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

n= 103, number of events= 75

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

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

Concordance= 0.653 (se = 0.032 )

Likelihood ratio test= 17.63 on 3 df, p=5e-04

Wald test = 15.76 on 3 df, p=0.001

Score (logrank) test = 16.71 on 3 df, p=8e-04

```
tbl_regression(coxfit, exponentiate = TRUE,)
```

Characteristic	HR <sup>1</sup>	95% CI <sup>1</sup>	p-value
year	0.89	0.78, 1.01	0.076

age	1.03	1.00, 1.06	0.029
surgery	0.37	0.16, 0.88	0.024

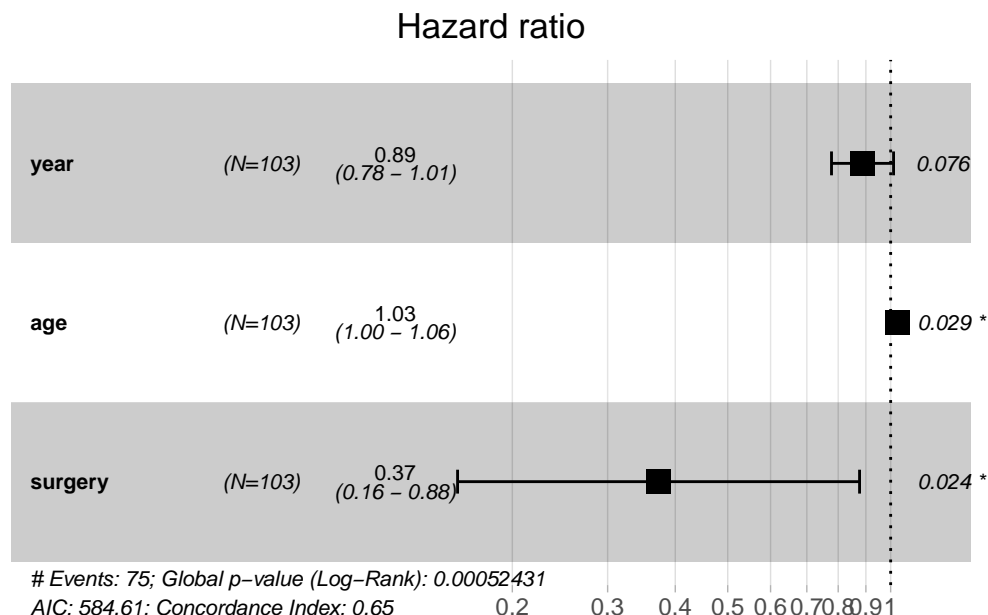
<sup>1</sup>HR = Hazard Ratio, CI = Confidence Interval

L'output des résultats reporte le logarithme des Risques Ratios (coef) ainsi que les RR ( $\exp(\text{coef})$ ). 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 d'un modèle probabiliste standard).

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

```
ggforest(coxfit)
```

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



## 15.3.2 Hypothèse PH

### 15.3.2.1 Test Grambsch-Therneau

#### Résidus de Schoenfeld

Traditionnellement, on utilise la fonction **cox.zph**.

Depuis la v3 du package (2020), il permet d'effectuer le test original de Grambsch-Therneau qui repose sur le calcul exact des résidus. Malheureusement, celui ci pose de gros problèmes en présence de covariables corrélées, même faiblement. Situation classique dans les sciences sociales. **Je ne déconseille fortement de l'utiliser**. Il est donc préférable de rester sur le test reposant sur les moindres carrés

ordinaires, implémenté jusqu'en 2023, et le seule disponible avec les autres outils: Stata, Python et Sas **†**. On maintiendra donc une reproductibilité du test dans le temps et dans l'espace des logiciel.

Le test peut utiliser plusieurs transformation de la durée. Par défaut la fonction utilise  $1 - KM$ , soit le complémentaire de l'estimateur de Kaplan-Meier (option `transform="km"`). Cette expression complémentaire permet juste d'avoir une suite de valeur partant de 0 (la valeur de la fonction de survie partant par définition de 1).

Test GLS (v3 de survival)... WARNING

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

Remarque: avec la v3 de survival, quelques options ont été ajoutées tel que `terms` qui permet pour une variable catégorielle à plus de deux modalités de choisir entre un sous test multiple sur la variable ( $k$  modalités  $\Rightarrow k-1$  degré de liberté) et une série de tests à 1 degré de liberté sur chaque modalité ( $k-1$  tests). De mon point de vue préférer la seconde solution avec `terms=FALSE`. Le test de Grambsch-Therneau sous sa forme multiple étant particulièrement sensible au nombre de degrés de liberté, il est à mon sens préférable d'évaluer la proportionnalité variable par variable, donc degré de liberté par degré de liberté.

Test OLS (v2 de survival - Stata - Sas - Python)...Use it

---

#### Listing 15.4 Récupération du test ols

---

```
source("https://raw.githubusercontent.com/mthevenin/analyse_duree/main/cox.zphold/cox.zphold.R")
```

---

Warning in is.R(): 'is.R' est obsolète.  
Voir `help("Deprecated")` et `help("base-deprecated")`.

---

**Listing 15.5** Exécution du test ols

---

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

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

On voit ici clairement que le test exact accentue la déviation vers la. C'est du tout simplement à la corrélation entre la variable *surgery* et la variable *year*. Les conclusions de S.Metzger sont ici bien vérifiées.

### 15.3.2.2 Introduction d'une interaction

Lorsque la covariable n'est pas continue, elle doit être impérativement transformée en indicatrice <sup>1</sup>. Penser à vérifier en amont que les résultats du modèle sont bien identiques avec le modèle estimé précédemment (ne pas oublier d'omettre le niveau en référence).

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(surgery),
      data = trans, tt = function(x, t, ...) x * t)
```

```
n= 103, number of events= 75
```

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

	exp(coef)	exp(-coef)	lower .95	upper .95
year	0.8842	1.1310	0.77564	1.0080
age	1.0293	0.9715	1.00253	1.0568

---

<sup>1</sup>c'est le cas ici, la variable *surgery* est bien codée (0;1)



```
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
```

```
tbl_regression(coxfit2, exponentiate = TRUE, estimate_fun = purrr::partial(style_ratio, digit
```

Characteristic	HR <sup>1</sup>	95% CI <sup>1</sup>	p-value
year	0.884	0.776, 1.008	0.066
age	1.029	1.003, 1.057	0.032
surgery	0.173	0.046, 0.649	0.009
tt(surgery)	1.002	1.000, 1.004	0.043

<sup>1</sup>HR = Hazard Ratio, CI = Confidence Interval

**Rappel:** le paramètre estimé pour **tt(surgery)** ne reporte pas un rapport de risques, mais un rapport de de deux rapports de risques. C'est bien une double différence sur l'échelle d'estimation (log).

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

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

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

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

```
[1] 1 2 3 5 6 8 9 12 16 17 18 21 28 30 32
[16] 35 36 37 39 40 43 45 50 51 53 58 61 66 68 69
[31] 72 77 78 80 81 85 90 96 100 102 110 149 153 165 186
[46] 188 207 219 263 285 308 334 340 342 583 675 733 852 979 995
[61] 1032 1386
```

Etape 2: allonger la base aux durées d'évènement

```
tvc = survSplit(data = trans, cut = cut, end = "stime", start = "stime0", event = "died")  
head(tvc, n=20 )
```

	id	year	age	surgery	transplant	wait	mois	compet	arm	stime0	stime	died
1	15	68	53	0	0	0	1	1	0	0	1	1
2	43	70	43	0	0	0	1	1	0	0	1	0
3	43	70	43	0	0	0	1	1	0	1	2	1
4	61	71	52	0	0	0	1	1	0	0	1	0
5	61	71	52	0	0	0	1	1	0	1	2	1
6	75	72	52	0	0	0	1	1	0	0	1	0
7	75	72	52	0	0	0	1	1	0	1	2	1
8	6	68	54	0	0	0	1	2	0	0	1	0
9	6	68	54	0	0	0	1	2	0	1	2	0
10	6	68	54	0	0	0	1	2	0	2	3	1
11	42	70	36	0	0	0	1	1	0	0	1	0
12	42	70	36	0	0	0	1	1	0	1	2	0
13	42	70	36	0	0	0	1	1	0	2	3	1
14	54	71	47	0	0	0	1	1	0	0	1	0
15	54	71	47	0	0	0	1	1	0	1	2	0
16	54	71	47	0	0	0	1	1	0	2	3	1
17	38	70	41	0	1	5	1	1	0	0	1	0
18	38	70	41	0	1	5	1	1	0	1	2	0
19	38	70	41	0	1	5	1	1	0	2	3	0
20	38	70	41	0	1	5	1	1	0	3	5	1

On vérifie qu'on obtient les même résultats avec le modèle sans tv

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

Call:

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

	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: on génère la variable dynamique de sorte que les personnes n'apparaissent pas greffés avant l'opération

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

## Estimation du modèle

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

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

summary(tvcsfit)
```

Call:

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

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
tvcs	-0.08221	0.92108	0.30484	-0.270	0.7874

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

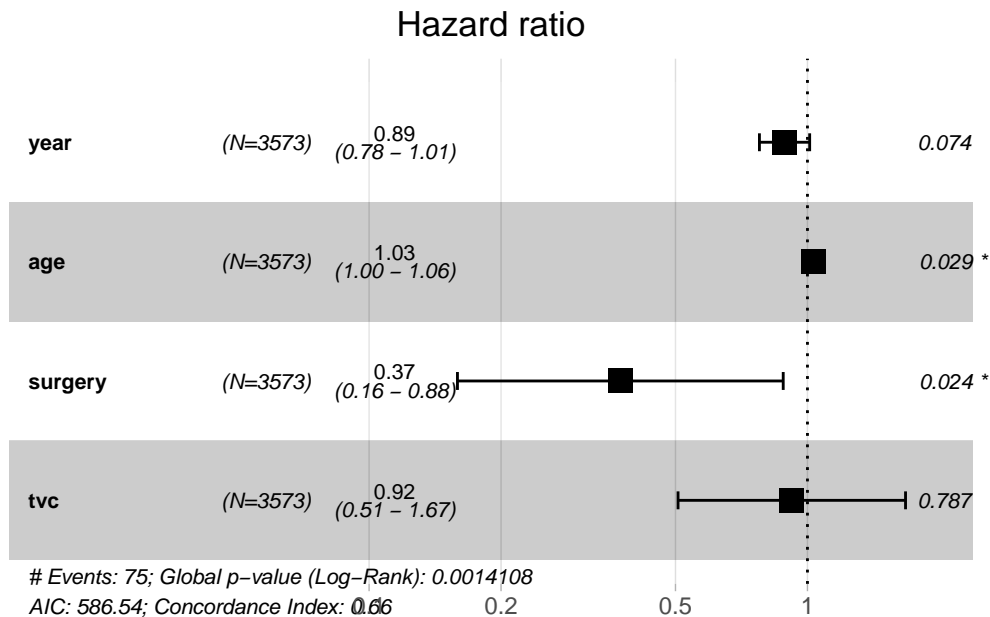
```
tbl_regression(tvcsfit, exponentiate = TRUE, estimate_fun = purrr::partial(style_ratio, digits = 3))
```

Characteristic	HR <sup>1</sup>	95% CI <sup>1</sup>	p-value
year	0.887	0.777, 1.012	0.074
age	1.031	1.003, 1.059	0.029
surgery	0.374	0.159, 0.880	0.024
tvcs	0.921	0.507, 1.674	0.8

<sup>1</sup>HR = Hazard Ratio, CI = Confidence Interval

```
ggforest(tvcfit)
```

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



## 15.4 Modèles à durée discrète

Pour la durée, on va utiliser la variable *mois* (regroupement sur 30 jours de *stime*).

La fonction **uncount** du package **tidyr** permettra de splitter la base aux durées d'observation. C'est ici la principale différence avec le modèle de Cox qui est une estimation aux durées d'évènement

```
trans <- read.csv("https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases/trans.csv")
```

La variable *mois*, va être supprimée avec **uncount**. Comme on en aura besoin plus loin pour générer proprement la variable évènement, on peut créer ici une variable miroir.

```
trans$T = trans$mois
```

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

```
head(dt,11)
```

	id	year	age	died	stime	surgery	transplant	wait	compet	T
48	1	67	30	1	50	0	0	0	1	2
49	1	67	30	1	50	0	0	0	1	2
10	2	68	51	1	6	0	0	0	1	1
18	3	68	54	1	16	0	1	1	1	1
36	4	68	40	1	39	0	1	36	2	2
37	4	68	40	1	39	0	1	36	2	2
20	5	68	20	1	18	0	0	0	1	1
5	6	68	54	1	3	0	0	0	2	1
466	7	68	50	1	675	0	1	51	1	23
467	7	68	50	1	675	0	1	51	1	23
468	7	68	50	1	675	0	1	51	1	23

On va générer une variable type compteur pour mesurer la durée à chaque point d'observation.

```
dt$x=1
dt$t = ave(dt$x,dt$id, FUN=cumsum)
head(dt, n=8)
```

	id	year	age	died	stime	surgery	transplant	wait	compet	T	x	t
48	1	67	30	1	50	0	0	0	1	2	1	1
49	1	67	30	1	50	0	0	0	1	2	1	2
10	2	68	51	1	6	0	0	0	1	1	1	1
18	3	68	54	1	16	0	1	1	1	1	1	1
36	4	68	40	1	39	0	1	36	2	2	1	1
37	4	68	40	1	39	0	1	36	2	2	1	2
20	5	68	20	1	18	0	0	0	1	1	1	1
5	6	68	54	1	3	0	0	0	2	1	1	1

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 = arrange(dt,id,t)
dt$died[dt$t<dt$T]=0
head(dt, n=8)
```

	id	year	age	died	stime	surgery	transplant	wait	compet	T	x	t
1	1	67	30	0	50	0	0	0	1	2	1	1
2	1	67	30	1	50	0	0	0	1	2	1	2
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	0	39	0	1	36	2	2	1	1

6	4	68	40	1	39	0	1	36	2	2	1	2
7	5	68	20	1	18	0	0	0	1	1	1	1
8	6	68	54	1	3	0	0	0	2	1	1	1

### 15.4.1 $f(t)$ quantitative

Avec un effet quadratique d'ordre 3 <sup>^</sup>[Attention ici cela marche bien. Bien vérifier qu'il n'y a pas un problème d'overfitting, comme c'est le cas dans le TP.

On centre également les variables *year* et *age* sur leur valeur moyenne pour donner un sens à la constante

```
dt$t2=dt$t^2
dt$t3=dt$t^3

my = mean(dt$year)
dt$yearb = dt$year - my
ma = mean(dt$age)
dt$ageb = dt$age - ma

dtfit = glm(died ~ t + t2 + t3 + yearb + ageb + surgery, data=dt, family="binomial")
summ(dtfit, confint=TRUE, exp=TRUE)
```

Observations	1127
Dependent variable	died
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(6)$	90.69
p	0.00
Pseudo-R <sup>2</sup> (Cragg-Uhler)	0.20
Pseudo-R <sup>2</sup> (McFadden)	0.16
AIC	474.67
BIC	509.86

```
tbl_regression(dtfit, exponentiate = TRUE, estimate_fun = purrr::partial(style_ratio, digits
```

Characteristic	OR <sup>†</sup>	95% CI <sup>†</sup>	p-value
t	0.689	0.582, 0.805	<0.001

t2	1.014	1.005, 1.025	0.005
t3	1.000	1.000, 1.000	0.035
yearb	0.876	0.756, 1.011	0.072
ageb	1.034	1.006, 1.066	0.023
surgery	0.364	0.136, 0.815	0.024

<sup>1</sup>OR = Odds Ratio, CI = Confidence Interval

## 15.4.2 $f(t)$ en indicatrices

On va créer une variable de type discrète regroupant la variable  $t$  sur ses quartiles (pour l'exemple seulement, tous types de regroupement est envisageable).

On utilisera 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 dans la strate.

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

## Estimation du modèle

```
fit = glm(died ~ ct4 + yearb + ageb + surgery, data=dt2, family=binomial)
summ(fit, confint=TRUE, exp=TRUE)
```

```
tbl_regression(fit, exponentiate = TRUE, estimate_fun = purrr::partial(style_ratio, digits =
```

Characteristic	OR <sup>1</sup>	95% CI <sup>1</sup>	p-value
ct4			
[1,4]	—	—	
(4,11]	0.356	0.152, 0.792	0.014

	(11,23]	0.199	0.061, 0.541	0.003
	(23,60]	0.619	0.183, 1.981	0.4
yearb		0.816	0.677, 0.977	0.029
ageb		1.048	1.012, 1.089	0.011
surgery		0.330	0.113, 0.837	0.027

<sup>1</sup>OR = Odds Ratio, CI = Confidence Interval

## 15.5 Modèles paramétriques usuels

Pour le modèle de **Weibull** par exemple.

- De type **AFT**

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

```
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

```
tbl_regression(weibull, exponentiate = TRUE, estimate_fun = purrr::partial(style_ratio, digits=2))
```

Warning: The `exponentiate` argument is not supported in the `tidy()` method for `survreg` objects and will be ignored.



	exp(Est.)	2.5%	97.5%	z val.	p
(Intercept)	0.44	0.27	0.72	-3.29	0.00
t	0.69	0.59	0.81	-4.52	0.00
t2	1.01	1.00	1.02	2.83	0.00
t3	1.00	1.00	1.00	-2.11	0.03
yearb	0.88	0.76	1.01	-1.80	0.07
ageb	1.03	1.00	1.06	2.27	0.02
surgery	0.36	0.15	0.88	-2.25	0.02

Standard errors: MLE

Observations	197
Dependent variable	died
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(6)$	39.30
p	0.00
Pseudo-R <sup>2</sup> (Cragg-Uhler)	0.25
Pseudo-R <sup>2</sup> (McFadden)	0.15
AIC	236.48
BIC	259.46

	exp(Est.)	2.5%	97.5%	z val.	p
(Intercept)	1.17	0.77	1.79	0.73	0.47
ct4(4,11]	0.36	0.16	0.81	-2.47	0.01
ct4(11,23]	0.20	0.07	0.58	-2.96	0.00
ct4(23,60]	0.62	0.19	2.01	-0.80	0.42
yearb	0.82	0.68	0.98	-2.18	0.03
ageb	1.05	1.01	1.09	2.53	0.01
surgery	0.33	0.12	0.88	-2.21	0.03

Standard errors: MLE

Characteristic	exp(Beta)	95% CI <sup>1</sup>	p-value
year	0.162	-0.077, 0.401	0.2
age	-0.062	-0.110, -0.013	0.013
surgery	1.970	0.443, 3.498	0.011

<sup>1</sup>CI = Confidence Interval

- De type **PH**

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`. La syntaxe est quasiment identique.

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

```
weibullph = flexsurvreg(formula = Surv(stime, died) ~ year + age + surgery, data = trans, dist = "weibullPH")
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	4.21e-04	6.85e+04	2.59e+01	NA
year	7.06e+01		-9.01e-02	-2.20e-01	4.00e-02	6.64e-02	9.14e-01
age	4.46e+01		3.42e-02	7.11e-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	8.02e-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

## 15.6 Risques concurrents

Le package `cmprsk` pour l'analyse non paramétrique et le modèle de Fine-Gray (non traité).

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 `cell-ci`.

```
compet <- read.csv("https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases/tr
# variable compet
table(compet$compet)
```

```
0 1 2
28 56 19
```

```
# variable died
table(compet$died)
```

```
0 1
28 75
```

### 15.6.0.1 Incidences cumulées

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

*Pas de comparaison de groupes*

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

Estimates and Variances:

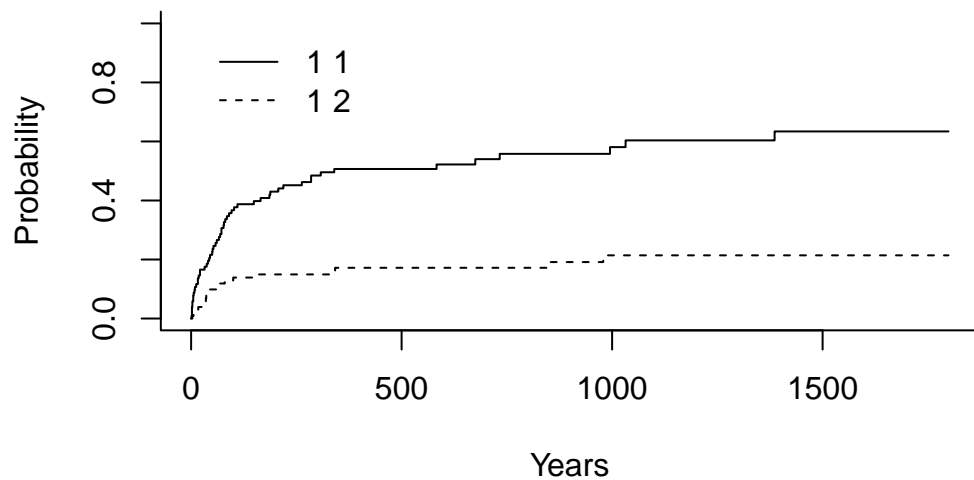
\$est

		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

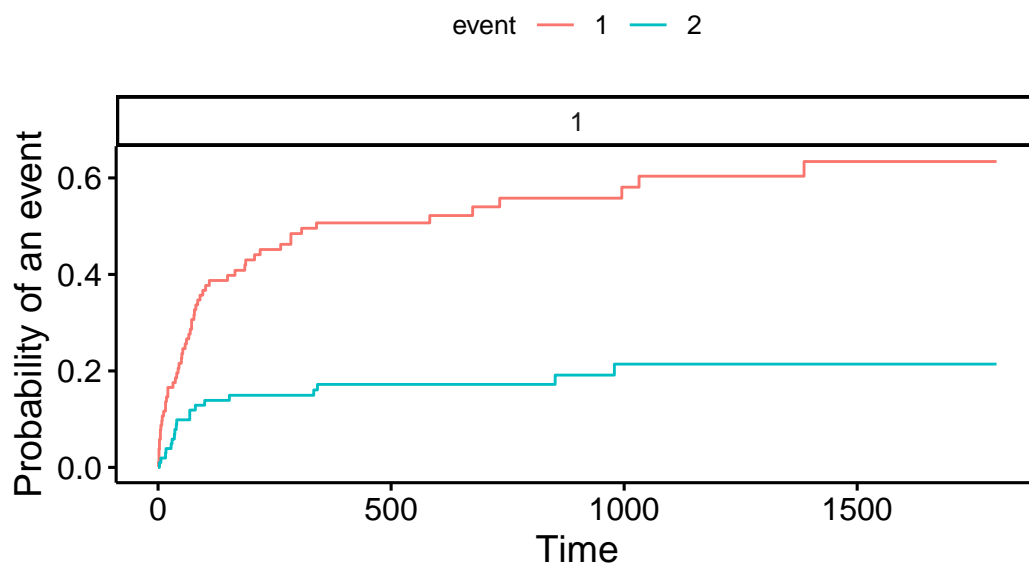
```
plot(ic)
```



Avec survminer

```
ggcompetingrisks(fit = ic)
```

## Cumulative incidence functions



*Comparaison de groupes*

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:

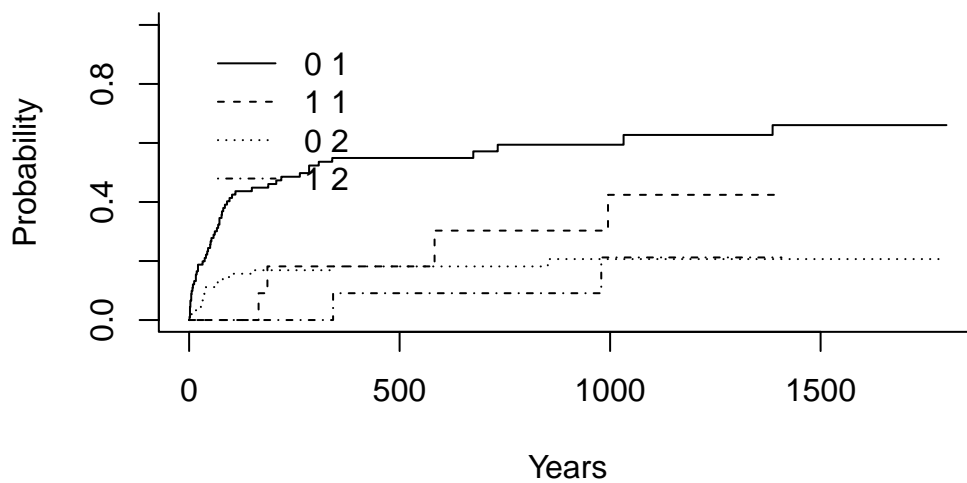
\$est

		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

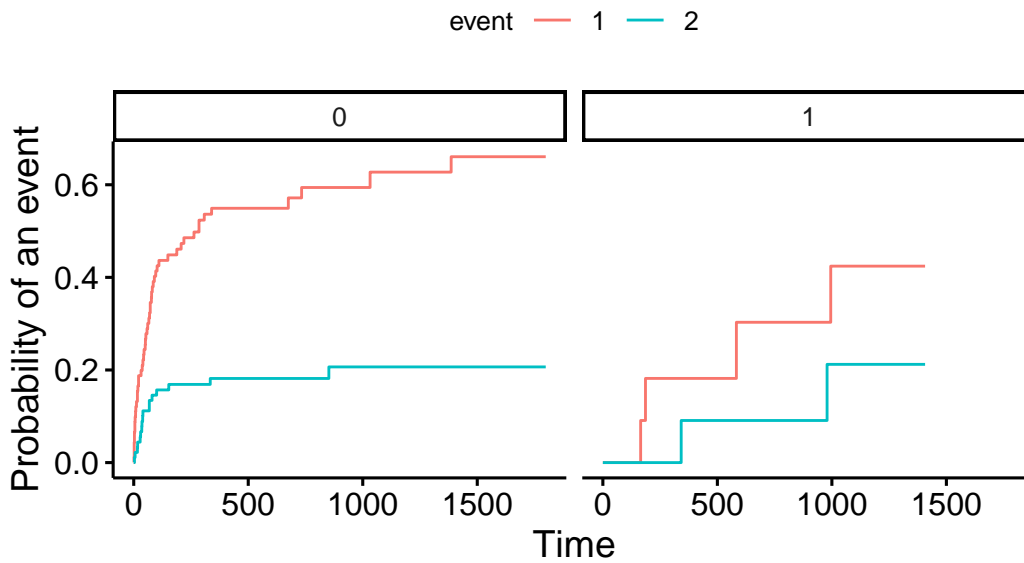
```
plot(ic)
```



Avec `survminer`, pour obtenir un seul graphique pour toutes les courbes ajouter l'option `multiple_panels = F`

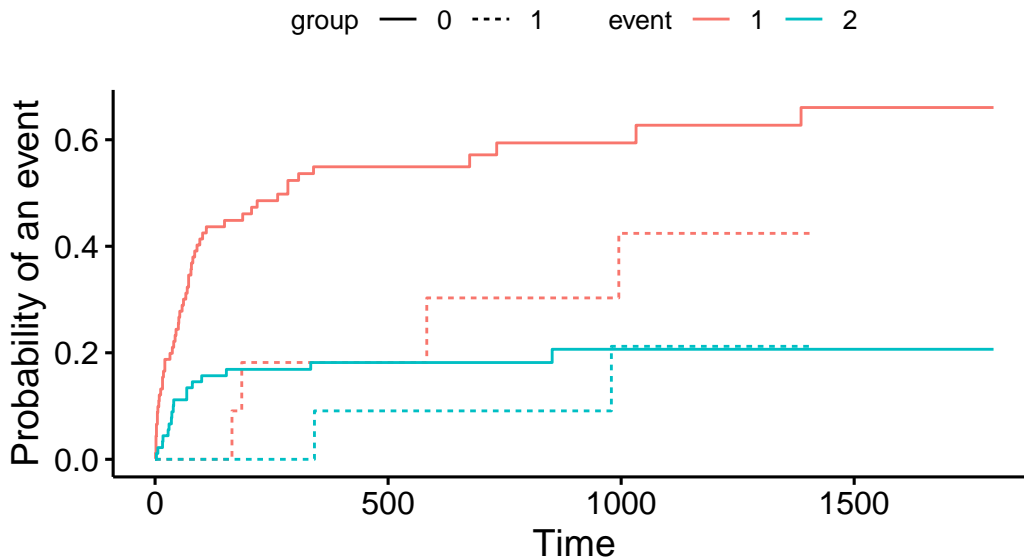
```
ggcompetingrisks(fit = ic)
```

## Cumulative incidence functions



```
ggcompetingrisks(fit = ic, multiple_panels = F)
```

## Cumulative incidence functions



### 15.6.0.2 Modèles

On va utiliser seulement le modèle multinomial à durée discrète, le modèle *fine-gray* pendant du modèle de Cox pour les risques concurrents étant fortement critiqué. Si une analyse de type *cause-specific* est envisageable (issues concurrentes traitées comme des censures à droites) on utilise simplement la fonction `coxph` de `survival`.

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

*Mise en forme de la base*

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

compet$T = compet$mois
td = uncount(compet, mois)
td = arrange(td, id)

td$x=1
td$t = ave(td$x, td$id, FUN=cumsum)
td$t2 = td$t^2

my = mean(td$year)
td$yearb = td$year - my
ma = mean(td$age)
td$ageb = td$age - ma

td$e = ifelse(td$t<td$T,0, td$compet)
```

*Estimation*

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

```
competfit = multinom(formula = e ~ t + t2 + yearb + ageb + surgery, data = td)
```

```
# weights:  21 (12 variable)
initial  value 1238.136049
iter   10 value 608.949443
iter   20 value 341.102661
iter   30 value 277.143136
iter   40 value 275.005451
final   value 275.005419
converged
```

```
tbl_regression(competfit, exponentiate = TRUE,)
```

Characteristic	OR <sup>1</sup>	95% CI <sup>1</sup>	p-value
1			
t	0.82	0.75, 0.88	<0.001

t2	1.00	1.00, 1.00	<0.001
yearb	0.88	0.75, 1.03	0.12
ageb	1.04	1.01, 1.08	0.012
surgery	0.32	0.11, 0.91	0.033
<hr/>			
2			
<hr/>			
t	0.82	0.71, 0.94	0.003
t2	1.00	1.00, 1.01	0.052
yearb	0.82	0.62, 1.07	0.14
ageb	1.01	0.96, 1.06	0.7
surgery	0.54	0.12, 2.50	0.4
<hr/>			

<sup>1</sup>OR = Odds Ratio, CI = Confidence Interval