NNT/NL : 2021AIXM0116/020ED62

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 30 mars 2021 par

François Lefebvre

Comparaison des modèles à risques instantanés multiplicatifs et additifs

Discipline
Biologie-Santé

Spécialité
Recherche Clinique et Santé Publique

École doctorale
ED62 Sciences de la Vie et de la Santé

Laboratoire
UMR1252 - SESSTIM (Sciences Economiques Sociales de la Santé Traitement de l'Information Médicale

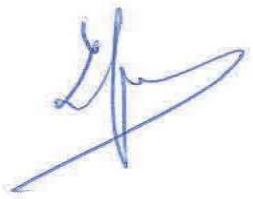
Composition du jury

Catherine Quantin PU-PH, HDR, Université Bourgogne Franche-Comté	Rapporteur
Matthieu Resche-Rigon PU-PH, HDR, Université de Paris	Rapporteur
Érik-André Sauleau PU-PH, HDR, Université de Strasbourg	Examinateur
Roch Giorgi PU-PH, HDR, Aix Marseille Université	Directeur de thèse

Je soussigné, François Lefebvre, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Roch Giorgi, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Molsheim le 31 janvier 2021



Résumé

Dans le domaine biomédical, l'étude des covariables associées à la survie est le plus souvent réalisée à l'aide du modèle de Cox. Avec ce modèle à risques instantanés multiplicatifs, les covariables sont supposées agir de manière multiplicative sur le risque instantané de base qui est une fonction non-paramétrique dépendant du temps. De plus, les effets des covariables sont supposés être constants au cours du temps, correspondant à l'hypothèse de proportionnalité des risques instantanés. Quand cette hypothèse n'est pas vérifiée, il faut utiliser soit une extension du modèle de Cox, soit un autre type de modèle. C'est dans ce contexte qu'Aalen a proposé un modèle à risques instantanés additifs. Dans ce modèle non-paramétrique, l'effet des covariables est modélisé par des fonctions de régression agissant de manière additive sur le risque instantané de base. Celui-ci, tout comme les fonctions de régression des covariables, sont non-paramétriques et peuvent varier dans le temps. Un avantage de ce modèle est que les fonctions de régression ne nécessitent pas d'hypothèses à l'exception de la linéarité des effets des covariables et permettent de mesurer l'effet des covariables au cours du temps. Toutefois, ces fonctions sont plus complexes à interpréter que les paramètres du modèle de Cox. Par conséquent, ce modèle est faiblement utilisé, ses fonctions représentant l'augmentation du risque instantané due aux covariables à la différence de l'exponentielle de chaque paramètre du modèle de Cox qui s'interprète comme un risque relatif. Plus récemment des modèles permettant de prendre en compte simultanément des covariables dont certaines ont des effets additifs et d'autres des effets multiplicatifs ont été développés.

Pour modéliser correctement les covariables, quand le type d'effet des covariables (additif ou multiplicatif) sur le risque instantané de base est inconnu, un certain nombre d'outils ont été développés. Ils permettent d'aider à la modélisation et de vérifier la bonne adéquation d'un modèle à risques instantanés multiplicatifs ou additifs aux données. Toutefois, il n'existe pas de stratégie permettant de modéliser correctement des données de survie selon que l'on souhaite utiliser un modèle à risques instantanés multiplicatifs ou additifs. Le premier objectif de ce travail fut de proposer une stratégie de modélisation des données de survie avec un modèle à risques instantanés multiplicatifs ou additifs en utilisant les différents outils diagnostiques fréquemment utilisés et d'autres moins connus. Cette stratégie a été appliquée à des données de survie issues d'une cohorte de patients ayant présenté un infarctus du myocarde et a permis de montrer que l'on obtenait des modèles s'ajustant correctement aux données dont les conclusions étaient similaires mais avec des différences en termes d'interprétation.

Une fois les deux modèles obtenus, il est intéressant de savoir lequel s'ajuste le mieux

aux données. Le second objectif consista à développer une méthode permettant de sélectionner parmi les deux types de modèles celui qui est le plus approprié pour une base de données particulière. Pour cela, une approche basée sur les pseudo-résidus qui sont, pour chaque sujet, la différence entre une estimation non-paramétrique de la survie (une pseudo-observation) et une estimation de la survie obtenue en utilisant un modèle de régression a été développée. L'utilisation de la somme des carrés des pseudo-résidus comme outil permettant de sélectionner le modèle le plus approprié a été proposée dans le cadre univarié puis dans le cadre multivarié en y incluant les modèles prenant en compte simultanément des effets multiplicatifs et additifs. Les performances de cette approche ont été étudiées par simulations et par applications sur des données réelles. Ce travail propose ainsi de nouveaux outils pouvant aider les biostatisticiens à réaliser un modèle à risques instantanés multiplicatifs ou additifs et à retenir celui qui est le plus approprié.

Mots clés : modèle de Cox, modèle de Aalen, pseudo-observations, pseudo-résidus

Abstract

In biostatistics, the study of covariates associated with survival is most often carried out using the Cox model. With this multiplicative hazards regression model, the covariates are assumed to act in a multiplicative manner on the baseline hazard which is a non-parametric time-dependent function. Moreover, the effects of the covariates are assumed to be constant over time, corresponding to the hypothesis of proportionality of the hazards. When this hypothesis is not verified, either an extension of the Cox model or another type of model must be used. In this context, Aalen proposed an additive hazards regression model. In this non-parametric model, the effect of the covariates is modelled by regression functions acting additively on the baseline hazard. The latter, like the covariate regression functions, are non-parametric and can vary over time. An advantage of this model is that the regression functions do not require assumptions except for the linearity of the effects of the covariates and allow the effect of the covariates to be measured over time. However, these functions are more complex to interpret than the parameters of the Cox model. Consequently, the use of this model is low: the functions represent the increase in hazard due to the covariates as opposed to the exponential of each parameter of the Cox model which is interpreted as a relative risk. More recently, hazards regression models have been developed which simultaneously take into account covariates, some of which have additive effects and others multiplicative effects.

In order to model the covariates correctly, when the type of effect of the covariates (additive or multiplicative) on the baseline hazard is unknown, a number of tools have been developed. They can be used to assist in modelling and to check the goodness of fit of multiplicative or additive hazards regression models. However, there is no strategy for correctly modelling survival data depending on whether one wishes to use a multiplicative or additive hazards regression model. The first objective of this work was to propose a strategy for modelling survival data with a multiplicative or additive hazards regression model using the different diagnostic tools frequently used and others less known. This strategy was applied to survival data from a cohort of patients with myocardial infarction and showed that obtained models fitted correctly the data with similar conclusions but with differences in interpretation.

Once both models are obtained, it is interesting to know which one fits the data best. The second objective was to develop a method to select from the two types of models the one that is most appropriate for a particular database. For this, an approach based on pseudo-residuals, which are, for each subject, the difference between a non-parametric estimate of survival (a pseudo-observation) and an estimate of survival obtained using a regression model, was developed. The use of the sum of squares of

pseudo-residuals as a tool for selecting the most appropriate model was proposed in the univariate framework and then in the multivariate framework by including models that simultaneously take into account multiplicative and additive effects. The performance of this approach has been studied by simulations and applications on real data. Thus this work proposes new tools that can help biostatisticians to perform multiplicative or additive hazards regression models and to select the most appropriate one.

Keywords: Cox model, Aalen's model, pseudo-observations, pseudo-residuals

Remerciements

Ce travail de thèse est l'aboutissement d'un travail personnel mais aussi collectif. En effet, il n'aurait pas pu voir le jour sans l'intervention, le soutien, l'accompagnement et la réflexion de nombreux acteurs. Je voudrais en premier remercier mon directeur de thèse, Roch Giorgi, qui a su m'accompagner dans ce travail et prendre régulièrement le temps nécessaire afin de résoudre les différents problèmes que nous avons rencontrés. La rigueur qu'il a essayée de me transmettre portera, je l'espère, de nombreux fruits. Il est rare en effet, que l'on soit trop précis. Je tenais aussi à remercier Nicolas Meyer de m'avoir permis de faire cette recherche. Je remercie les membres de mon jury d'avoir accepté de lire et de juger ce travail. Je remercie les professeurs Catherine Quantin et Matthieu Resche-Rigon d'en être les rapporteurs et le professeur Érik-André Sauleau d'en être un examinateur.

Durant ces années de thèse, j'ai pu profiter de l'aide et de l'expertise, souvent à distance, mais aussi parfois de visu, de nombreux collègues. Je pense notamment aux membres du groupe CENSUR, aux post-doctorants, Célia et Nathalie, aux doctorants, notamment Juste et Darlin, et aux étudiants de master, Casimir, Anna, Nicolas et Pierre. Qu'ils en soient remerciés. Je voudrais aussi remercier mes collègues strasbourgeois, notamment François Séverac et Nicolas Tuzin. Enfin, j'aimerais remercier ma famille, mes parents, ma femme et mes enfants et mes frère et sœur de leur soutien. Et pour terminer, je remercie le professeur Odd Olai Aalen pour avoir développé son modèle si fascinant sans qui ce travail n'aurait pu voir le jour.

Table des matières

Résumé 4

Abstract 6

Remerciements 7

Table des matières 8

Table des figures 10

Liste des tableaux 12

Introduction 14

1 Modèles de survie 15
- 1.1 Généralités sur la survie . . . 16
  - 1.1.1 Fonctions de survie . . . 16
  - 1.1.2 Concept de la censure . . . 17
  - 1.1.3 Processus de comptage . . . 17
  - 1.1.4 Estimateurs non-paramétriques . . . 18
- 1.2 Modèles à risques instantanés multiplicatifs . . . 19
  - 1.2.1 Modèle de Cox . . . 19
  - 1.2.2 Extensions du modèle de Cox . . . 22
- 1.3 Modèles additifs . . . 24
  - 1.3.1 Modèle d'Aalen . . . 24
  - 1.3.2 Extension du modèle d'Aalen . . . 27
  - 1.3.3 Autres modèles additifs . . . 27
- 1.4 Modèles multiplicatifs et additifs . . . 28
  - 1.4.1 Modèle de Cox-Aalen . . . 28
  - 1.4.2 Autres modèles multiplicatifs et additifs . . . 28

2 Modélisation du risque instantané 30
- 2.1 Outils diagnostiques . . . 30
  - 2.1.1 Pseudo-observations . . . 30
  - 2.1.2 Graphiques d'Arjas . . . 32
  - 2.1.3 Processus de résidus de martingale . . . 33

2.2 Proposition d'une stratégie de modélisation des modèles à risques instantanés multiplicatifs . . . 34
2.3 Proposition d'une stratégie de modélisation des modèles à risques instantanés additifs . . . 35
2.4 Résultats . . . 35
2.4.1 Modélisation multiplicative . . . 36
2.4.2 Modélisation additive . . . 45
2.5 Discussion . . . 53

3 Comparaison des modèles à risques instantanés multiplicatifs et additifs 57
3.1 Pseudo-résidus . . . 57
3.1.1 Définition . . . 58
3.1.2 Somme des carrés des pseudo-résidus . . . 58
3.2 Étude de simulations . . . 59
3.2.1 Une variable . . . 59
3.2.2 Deux variables . . . 59
3.3 Analyses des simulations . . . 60
3.3.1 Une seule variable continue . . . 61
3.3.2 Une seule variable binaire . . . 63
3.3.3 Deux covariables . . . 65
3.4 Applications . . . 74
3.4.1 Cirrhose biliaire primitive . . . 74
3.4.2 Cancer du sein . . . 75
3.4.3 Infarctus du myocarde . . . 76
3.5 Discussion . . . 76

Conclusion 82

Bibliographie 83

ANNEXES 88
A Relation entre la somme des carrés des pseudo-residus et l'estimation de Kaplan-Meier . . . 88
B Programme R . . . 91
C Figures annexes . . . 112

Table des figures

2.1 Résidus de martingale 36
2.2 Résidus de Schoenfeld 38
2.3 Pseudo-observations en fonction de l'âge. Modèle à risques instantanés multiplicatif 39
2.4 Graphiques d'Arjas pour l'âge 40
2.5 Graphiques d'Arjas pour l'insuffisance cardiaque 41
2.6 Graphiques d'Arjas pour la fibrillation ventriculaire 41
2.7 Résidus de Schoenfeld. Modèles multivariés 43
2.8 Rapport des risques instantanés pour l'âge 44
2.9 Rapport des risques instantanés l'insuffisance cardiaque, le diabète et la fibrillation ventriculaire 45
2.10 Pseudo-observations en fonction de l'âge. Modèle à risques instantanés additif 46
2.11 Processus des résidus de martingale 47
2.12 Processus des résidus de martingale 48
2.13 Graphiques d'Arjas ajustés sur l'âge. Modèle additif 49
2.14 Graphiques d'Arjas ajustés sur l'âge. Modèle additif 50
2.15 Processus des résidus de martingale 52
2.16 Processus des résidus de martingale 52
2.17 Risques instantanés cumulés de l'âge 53

3.1 Distributions des différences des SCPR. Une variable générée continue. 61
3.2 Distributions des différences des SCPR. Une variable générée binaire. . 64
3.3 Distributions des différences des SCPR. Deux variables générées identiquement. 66
3.4 Pourcentage de cas pour lesquels le meilleur modèle est retenu. 67
3.5 Distributions des différences des SCPR. Deux variables générées identiquement. 68
3.6 Distributions des différences des SCPR. Deux variables générées identiquement. 69
3.7 Distributions des différences des SCPR. Deux variables générées différemment. 70
3.8 Distributions des différences des SCPR. Deux variables générées différemment. 71
3.9 Distributions des différences des SCPR. Deux variables générées différemment. 72

3.10 Moyenne des pseudo-résidus. Deux variables générées identiquement. 73

.1 Résidus de martingale. 112
.2 Distributions des différences des SCPR. Une variable générée continue. 113
.3 Distributions des différences des SCPR. Une variable générée binaire. . 114
.4 Pourcentage de cas pour lesquels le meilleur modèle est retenu. 115
.5 Pourcentage de cas pour lesquels le meilleur modèle est retenu. 116

Liste des tableaux

2.1 Significativité des tests de corrélation 37
2.2 Significativité des tests de corrélation. Modèles multivariés 42
2.3 Paramètres $\hat{\beta}$ et rapports des risques instantanés ajustés 43
2.4 Résultats des tests du χ^2 48
2.5 Résultats des tests du χ^2 51
2.6 Significativité des fonctions de régression du modèle d'Aalen 51

3.1 Comparaison entre les modèles de Cox et d'Aalen. Une variable continue générée 62
3.2 Comparaison entre les modèles de Cox et de Lin. Une variable continue générée. 62
3.3 Comparaison entre les modèles de Cox et d'Aalen. Une variable binaire générée. 64
3.4 Comparaison entre les modèles de Cox et de Lin. Une variable binaire générée. 65
3.5 SCPR estimées 74
3.6 SCPR estimées 75
3.7 SCPR estimées 76

Introduction

L'étude de la survie est un domaine des statistiques qui s'est développé à l'origine grâce aux démographes, notamment à partir du XVII^{e} siècle. En effet, la première table de mortalité a été réalisée en 1662 par John Graunt (1620-1674) et l'on doit à William Petty (1623-1687) la notion d'espérance de vie à la naissance et d'espérance de vie résiduelle. En France, Antoine Deparcieux (1703-1768) écrivit son *Essai sur les probabilités de la durée de la vie humaine, d'où l'on déduit la manière de déterminer les rentes viagères tant simples que tantines, précédé d'une courte explication sur les rentes à terme, ou annuités, et accompagné d'un grand nombre de tables* en 1746 dans lequel on trouve les célèbres « Tables de Mortalité ». Cet ouvrage est encore aujourd'hui considéré comme l'ouvrage fondateur de la science actuarielle. C'est au XIX^{e} siècle qu'apparurent les tables de mortalité catégorisées par des variables telles que le sexe, la nationalité et la catégorie socio-professionnelle ainsi que les premières modélisations de la probabilité de décéder à un âge donné. Benjamin Gompertz (1779-1865) décrivit en 1825 une loi de probabilité modélisant le taux de mortalité.
Ce n'est qu'au XX^{e} siècle que les applications à la médecine furent proposées. PE. Böhmer en 1912 [11] proposa une méthode non-paramétrique d'estimation de la survie dont l'idée fut reprise et rendue célèbre par la publication de 1958 de Edward Kaplan et Paul Meier [23]. La formule de la variance fut proposée par Greenwood en 1926 [20]. Des tests de comparaison des courbes de survie furent développés, notamment le test du Logrank [40] puis dans un but de prendre en compte l'effet simultané de différentes variables, des modèles de survie basés sur la loi exponentielle puis sur la loi de Weibull furent proposés. En 1972, David Cox [15] proposa un modèle de survie semi-paramétrique modélisant directement le risque instantané. Celui-ci est le produit d'une fonction non-paramétrique caractérisant le risque instantané de base et de l'exponentielle du produit du vecteur des covariables par un vecteur de paramètres inconnus. Ce modèle a l'avantage d'une grande souplesse, en ne paramétrisant pas la fonction de risque instantané de base, et d'avoir une interprétation aisée, l'exponentielle des paramètres étant assimilable à des risques relatifs. Ses propriétés font que le rapport des risques instantanés est supposé constant au cours du temps et l'effet des covariables est supposé log-linéaire et multiplicatif sur le risque instantané de base.

Le développement de la théorie des processus stochastiques (théorie des martingales, des intégrales stochastiques et de la convergence en loi de processus) a permis dans un premier temps d'obtenir un estimateur du risque instantané cumulé, l'estimateur de Nelson-Aalen [37], [38] et [1], et un cadre général à l'ensemble des modèles de survie. Andersen et Gill ont, en 1982, écrit le modèle de Cox en termes de processus

stochastiques [6]. Puis les travaux d'Aalen lui ont permis de développer un nouveau modèle permettant de modéliser les effets des covariables de manière additive sur le risque instantané de base. Dans le premier chapitre, les différents modèles de survie modélisant le risque instantané de base avec un effet multiplicatif ou additif seront présentés avec les outils diagnostiques permettant de vérifier les hypothèses de ces modèles. Le deuxième chapitre sera consacré à la description d'une stratégie de modélisation d'un modèle de survie selon que l'on utilise un modèle à risques instantanés multiplicatifs ou additifs. Le troisième chapitre décrira une méthode permettant de sélectionner le meilleur modèle en se basant sur les pseudo-résidus.

1. Modèles de survie

Sommaire

1.1 Généralités sur la survie . . . 16
1.1.1 Fonctions de survie . . . 16
1.1.2 Concept de la censure . . . 17
1.1.3 Processus de comptage . . . 17
1.1.4 Estimateurs non-paramétriques . . . 18
1.1.4.1 Estimateur de Kaplan-Meier . . . 18
1.1.4.2 Estimateur de Nelson-Aalen . . . 18
1.2 Modèles à risques instantanés multiplicatifs . . . 19
1.2.1 Modèle de Cox . . . 19
1.2.1.1 Définition du modèle . . . 19
1.2.1.2 Estimation des paramètres du modèle . . . 20
1.2.1.3 Interprétation des paramètres . . . 20
1.2.1.4 Test de l'effet des paramètres . . . 21
1.2.2 Extensions du modèle de Cox . . . 22
1.2.2.1 Absence de propotionnalité des risques instantanés . . . 22
1.2.2.2 Absence de log-linéarité . . . 23
1.2.2.3 Absence de propotionnalité des risques instantanés et de log-linéarité . . . 23
1.3 Modèles additifs . . . 24
1.3.1 Modèle d'Aalen . . . 24
1.3.1.1 Définition du modèle . . . 24
1.3.1.2 Estimation des fonctions de régression du modèle . . . 24
1.3.1.3 Interprétation des fonctions de régression . . . 25
1.3.1.4 Test des fonctions de régression . . . 26
1.3.2 Extension du modèle d'Aalen . . . 27
1.3.3 Autres modèles additifs . . . 27
1.3.3.1 Modèle semi-paramétrique de Lin . . . 27
1.3.3.2 Modèle semi-paramétrique de McKeague et Sasieni . . . 27
1.4 Modèles multiplicatifs et additifs . . . 28
1.4.1 Modèle de Cox-Aalen . . . 28
1.4.2 Autres modèles multiplicatifs et additifs . . . 28

1.1. Généralités sur la survie

1.1.1. Fonctions de survie

L'analyse de survie s'intéresse à la durée T^* entre une date d'origine et la survenue d'un événement. À l'origine, il s'agissait du décès, d'où le fait que cette branche des statistiques s'appelle l'analyse de survie. Cette variable T^* est donc une variable quantitative continue positive, donc définie sur $\mathbb{R}+$. Dans certains cas, cette variable aléatoire T^* peut être analysée de manière discrète. Il existe plusieurs façons de décrire cette variable qui sont présentées ci-dessous. Comme pour toute variable aléatoire, on utilise la densité de probabilité, qui pour T^* est $f(t)$ et équivaut à :

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

La fonction de répartition $F(t)$ est la probabilité que l'événement survienne avant t et s'écrit :

$$F(t) = \int_0^t f(u)du = P(T^* < t).$$

En analyse de survie, on s'intéresse souvent à la probabilité de ne pas présenter l'événement avant t, qui est la fonction de survie. Elle s'écrit :

$$S(t) = P(T^* \geq t) = 1 - F(t).$$

Puisque $F(t)$ est une fonction monotone croissante, $S(t)$ est une fonction monotone décroissante telle que $S(0) = 1$ et $\lim_{t \to \infty} S(t) = 0$.

On utilise souvent une autre fonction, appelée risque instantané ou intensité et qui s'écrit :

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T^* < t + dt | T^* \geq t)}{dt}. \tag{1.1}$$

$\lambda(t)dt$ est la probabilité de présenter l'événement entre t et $t + dt$ sachant que le sujet n'a pas présenté l'événement avant t.

On a :

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = \frac{-dln(S(t))}{dt}$$

$$S(t) = e^{-\int_0^t \lambda(u)du}$$

et

$$f(t) = \lambda(t)e^{-\int_0^t \lambda(u)du}.$$

La fonction de risque cumulé $\Lambda(t)$ s'écrit :

$$\Lambda(t) = \int_0^t \lambda(u)du = -ln(S(t))$$

Chacune des fonctions ci-dessus peut être obtenue à partir d'une autre.

C'est la fonction de risque instantané $\lambda(t)dt$ qui est modélisée avec les modèles de survie à risques instantanés additifs et multiplicatifs, notamment avec le modèle de Cox et le modèle d'Aalen.

1.1.2. Concept de la censure

Les données de survie ont la particularité de ne pas toujours être complètes, c'est-à-dire que tous les sujets inclus dans l'étude n'ont pas toujours présenté à la date de point l'événement d'intérêt. Celle-ci correspond à la date à laquelle l'étude s'arrête et donc à laquelle on ne tient plus compte de l'information ultérieure. On dispose ainsi pour ces sujets d'une donnée partielle, à savoir un délai au cours duquel on sait que l'événement d'intérêt ne s'est pas produit. Dans ce cas, on parle de données censurées à droite. Il existe également des données censurées à gauche et des données censurées par intervalle. Dans ce travail, seules ont été étudiées les données censurées à droite. En analyse de survie, deux grands mécanismes sont à l'origine de la censure à droite. Soit l'événement n'a pas encore eu lieu à la date de point et les sujets sont exclus vivants, soit le sujet a quitté l'étude et il est perdu de vue. En présence de données censurées à droite, on n'observe pas pour tous les sujets le temps de survie T^* mais la durée T telle que $T = min(T^*, C)$ où C est le temps jusqu'à la censure. Par conséquent, on observe le couple (T, δ) avec T le temps d'observation du sujet et δ une indicatrice qui vaut 1 si le sujet présente l'événement au temps T ou 0 s'il est censuré.

1.1.3. Processus de comptage

Plutôt que de considérer la durée de survie T, il est possible de s'intéresser au processus de comptage $N(t)$ qui vaut 0 tant que l'événement ne s'est pas produit et 1 après, avec t le temps variant dans l'intervalle $(0, t_{max})$, avec souvent $t_{max} = \infty$. Analytiquement, on a : $N(t) = I(t \geq T^*), avec t \geq 0$ et où T^* est le temps de survie. Comme on l'a vu précédemment, quand on a des données censurées à droite, on n'observe pas le temps de survie T^* mais la durée T telle que $T = min(T^*, C)$ où C est le temps jusqu'à la censure. Le processus de comptage $N(t)$ s'écrit alors :

$$N(t) = I(t \geq T, \delta = 1)$$

et l'intensité du processus à l'instant t vaut :

$$\lambda(t) = Y(t)\alpha(t).$$

où $Y(t) = I(t \leq T)$ est l'indicateur de risque. La dérivée du processus de comptage $dN(t)$ vaut donc uniquement 1 à l'instant t correspondant à l'événement, 0 sinon. L'intensité cumulée $\Lambda(t)$ du processus de comptage $N(t)$, aussi appelée compensateur, vaut :

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t Y(u)\alpha(u)du.$$

La différence entre le processus de comptage $N(t)$ et son compensateur est une martingale :

$$M(t) = N(t) - \Lambda(t).$$

Ainsi, pour un échantillon de n sujets indexés par $i = 1, ..., n$, on a pour chaque sujet un indicateur de la présence du risque à l'instant t :

$$Y_i(t) = I(t \leq T_i)$$

et un processus de comptage de l'événement d'intérêt :

$$N_i(t) = I(t \geq T_i, \delta_i = 1).$$

Le nombre de sujets à risque à l'instant t est :

$$\overline{Y}(t) = \sum_{i=1}^{n} Y_i(t)$$

et le nombre d'événements à l'instant t est

$$\overline{N}(t) = \sum_{i=1}^{n} N_i(t).$$

1.1.4. Estimateurs non-paramétriques

1.1.4.1. Estimateur de Kaplan-Meier

L'estimateur qui maximise la vraisemblance de la fonction de survie est appelé l'estimateur de Kaplan-Meier. Il s'écrit de la manière suivante :

$$\hat{S}(t) = \prod_{\{i:t_i \leq t\}} \left(1 - \frac{d\overline{N}(t_i)}{\overline{Y}(t_i)}\right)$$

où $d\overline{N}(t_i)$ est le nombre d'événement à l'instant t_i et $\overline{Y}(t_i)$ le nombre de sujets à risque au temps t_i. L'estimateur de la variance de l'estimateur de Kaplan-Meier proposée par Greenwood [20] s'écrit :

$$\hat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{\{i:t_i \leq t\}} \frac{d\overline{N}(t_i)}{\overline{Y}(t_i)\left(\overline{Y}(t_i) - d\overline{N}(t_i)\right)}.$$

1.1.4.2. Estimateur de Nelson-Aalen

L'estimateur de Nelson-Aalen [37], [38] et [1] est un estimateur du risque instantané cumulé qui s'écrit :

$$\hat{H}(t) = \int_0^t \frac{d\overline{N}(s)}{\overline{Y}(s)}$$

Comme cette fonction ne comporte que des sauts, on peut également l'écrire :

$$\hat{H}(t) = \sum_{\{i:t_i \leq t\}} \frac{d\overline{N}(t_i)}{\overline{Y}(t_i)}.$$

Un estimateur de la variance de $\hat{H}(t)$ est $\hat{Var}\left(\hat{H}(t)\right) = \sum_{\{i:t_i \leq t\}} \frac{d\overline{N}(t_i)}{\overline{Y}^2(t_i)}$. L'estimateur de Nelson-Aalen est tel que l'on a la relation :

$$\sum_{i=1}^{n} \hat{H}(t_i) = \sum_{i=1}^{n} N(t_i).$$

1.2. Modèles à risques instantanés multiplicatifs

Les modèles à risques instantanés multiplicatifs supposent que le risque instantané est le produit entre le risque instantané de base et une fonction des covariables. Ils s'écrivent de la façon suivante :

$$\lambda(t|Z) = Y(t)\lambda_0(t)f(Z\beta) \tag{1.2}$$

où

- $Y(t)$ est la fonction indicatrice qui vaut 1 si l'événement n'a pas encore eu lieu à l'instant t, 0 sinon
- $\lambda_0(t)$ est la fonction du risque instantané de base qui dépend du temps
- f est une fonction qui dépend des covariables
- Z est une matrice de covariables de taille n (nombre de sujets) par r (nombre de covariables) avec pour chaque ligne i : $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ir})$
- β est le vecteur des r paramètres associés aux covariables du modèle
- T est le temps jusqu'à l'événement ou la censure

1.2.1. Modèle de Cox

1.2.1.1. Définition du modèle

Le modèle de Cox [15] est un modèle à risques instantanés multiplicatifs qui repose sur l'équation suivante :

$$\lambda(t) = Y(t)\lambda_0(t)e^{Z\beta} \tag{1.3}$$

où

- $\lambda_0(t)$ est la fonction de risque instantané de base. Elle correspond au risque instantané des sujets ayant leurs covariables fixées à la valeur de référence. C'est une fonction qui est non-paramétrique et qui dépend uniquement du temps.
- la fonction f est la fonction exponentielle qui ne dépend pas du temps, ce qui permet d'obtenir un risque intantané positif sans qu'il soit nécessaire d'ajouter une contrainte sur les valeurs des coefficients.

1.2.1.2. Estimation des paramètres du modèle

L'estimation du vecteur des paramètres β se fait par maximisation de la vraisemblance. Toutefois, afin de n'avoir pas à faire d'hypothèses sur la fonction $\lambda_0(t)$, Cox a proposé une vraisemblance partielle [16] qui ne dépend pas de $\lambda_0(t)$. Il a obtenu son résultat en utilisant des probabilités conditionnelles. Cette vraisemblance a été traduite en termes de processus de comptage par Andersen et Gill et c'est sous cette forme qu'elle est présentée [6].

Celle-ci vaut :

$$V_p(\beta) = \prod_i \prod_t \left(\frac{Y_i(t) e^{Z_i \beta}}{\sum_{j=1}^n Y_j(t) e^{Z_j \beta}} \right)^{dN_i(t)} . \tag{1.4}$$

Le logarithme de la vraisemblance partielle est donc :

$$LV_p(\beta) = \sum_i \sum_t \left[Y_i(t) Z_i \beta - ln \left(\sum_{j=1}^n Y_j(t) e^{Z_j \beta} \right) \right] dN_i(t). \tag{1.5}$$

On obtient un estimateur $\hat{\beta}$ de β comme solution du maximum du logarithme de la vraisemblance partielle à l'aide d'algorithmes tels que celui de Newton-Raphson. Le logarithme de la vraisemblance partielle LV_p est lié au logarithme de la vraisemblance totale LV_t par la relation :

$$LV_t = LV_p + \sum_t \left(\Delta \overline{N}(t) + ln(\Delta \overline{N}(t)) \times \Delta \overline{N}(t) \right)$$

et il a été montré qu'elle a des propriétés similaires [6] et [17], notamment que l'estimation $\hat{\beta}$ de β basée sur la maximisation de la vraisemblance partielle suit approximativement une loi normale $N(\beta, I^{-1})$ où I est la matrice d'information de Fisher de la vraisemblance partielle quand n tend vers l'infini. Une fois l'estimateur $\hat{\beta}$ obtenu, il est possible d'estimer le risque instantané de base que l'on obtient à l'aide de l'estimateur de Breslow $\hat{\lambda}_0(t, \beta) = \frac{d\overline{N}(t)}{\sum_{i=1}^n Y_i(t) e^{Z_i \beta}}$.

1.2.1.3. Interprétation des paramètres

Les estimations des paramètres β représentent le logarithme du surrisque associé à l'augmentation d'une unité de la variable Z. En présence d'une covariable binaire d'exposition Z_1 et de deux sujets ($i = 1, 2$) avec $Z_{11} = 1$ et $Z_{21} = 0$, on a alors pour le rapport de leurs risques instantanés :

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{e^{Z_{11}(t)\beta_1}}{e^{Z_{21}(t)\beta_1}} = e^{\beta_1}$$

qui représente le risque relatif de présenter un événement pour les sujets exposés par rapport aux sujets non-exposés. On remarque que ce rapport des risques instantanés ne dépend pas du temps. Cela correspond à l'hypothèse des risques instantanés

proportionnels.
En présence d'une covariable continue d'exposition Z_2 et de deux sujets ($i = 1,2$) avec $Z_{12} = Z_{22} + 1$, on a alors pour le rapport de leurs risques instantanés :

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{e^{Z_{12}(t)\beta_2}}{e^{Z_{22}(t)\beta_2}} = e^{\beta_2}$$

qui représente le risque relatif de présenter un événement pour les sujets dont la variable vaut une unité de plus. On remarque que ce rapport est constant quelle que soit la valeur de la variable Z_2. Cela correspond à l'hypothèse de log-linéarité.

1.2.1.4. Test de l'effet des paramètres

Afin d'étudier l'effet d'une covariable sur le risque instantané, trois différentes approches sont possibles. La statistique de Wald $(\hat{\beta} - \beta_0)' I(\beta)(\hat{\beta} - \beta_0)$ permet de tester l'hypothèse nulle H_0 selon laquelle $\beta = \beta_0$ tout comme la statistique du rapport de vraisemblance $-2ln\left(\frac{LV_p(\beta_0)}{LV_p(\hat{\beta})}\right)$ ou du score $U(\beta_0)' I(\beta_0)^{-1} U(\beta_0)$, avec $U(\beta_0) = \sum_{i=1}^{n} \int_0^\tau (X_i(t) - E(t,\beta_0)) dN_i(t)$. On a :

$$E(t,\beta) = \frac{S_1(t,\beta)}{S_0(t,\beta)}$$

$$I(t,\beta) = \sum_{i=1}^{n} \int_0^t \left(\frac{S_2(s,\beta)}{S_0(s,\beta)} - E(s,\beta)^{\otimes 2} \right) dN_i(s)$$

$$S_0(t,\beta) = \sum_{i=1}^{n} Y_i(t) e^{Z_i\beta}$$

$$S_1(t,\beta) = \sum_{i=1}^{n} Y_i(t) e^{Z_i\beta} Z_i'$$

$$S_2(t,\beta) = \sum_{i=1}^{n} Y_i(t) e^{Z_i\beta} Z_i'^{\otimes 2}.$$

Sous l'hypothèse nulle H_0, ces trois statistiques suivent une loi du χ^2 à r degrés de liberté correspondant au nombre total de paramètres et peuvent permettre de tester l'hypothèse selon laquelle $\beta = \beta_0$. Les trois tests obtenus sont asymptotiquement équivalents. Il peut être intéressant de ne pas tester tous les paramètres ensemble mais seulement certains. Avec le test de Wald, pour une seule variable, le test est équivalent à $\frac{\hat{\beta}}{\sqrt{\hat{var}(\hat{\beta})}}$ où $\hat{var}(\hat{\beta})$ est la variance de $\hat{\beta}$ estimée à partir de la matrice d'information de Fisher $I(\beta)$. Il est également possible de constituer des intervalles de confiance basée sur la statistique de Wald et qui s'écrivent : $IC_{95\%}(\beta) = \hat{\beta} \pm 1,96\sqrt{\hat{var}(\hat{\beta})}$.

1.2.2. Extensions du modèle de Cox

1.2.2.1. Absence de propotionnalité des risques instantanés

Le modèle de Cox suppose la proportionnalité des risques instantanés, c'est-à-dire la constance de l'effet au cours du temps. Si cette hypothèse n'est pas vérifiée, les estimations du modèle de Cox ne sont pas correctes et il est nécessaire d'utiliser une extension du modèle de Cox. L'extension la plus utilisée consiste à introduire des paramètres pouvant varier au cours du temps. Cette extension du modèle de Cox s'écrit :

$$\lambda(t) = Y(t)\lambda_0(t)e^{Z\beta(t)} \tag{1.6}$$

où $\beta(t)$ doit être modélisé soit sur la base de la connaissance de l'effet de la variable dans le temps, soit à l'aide des outils diagnostiques comme les résidus de Schoenfeld. Ceux-ci [44], aussi appelés résidus du score, leur somme sur tous les sujets étant égale à la statistique du score, sont définis pour les sujets présentant l'événement par :

$$r_s(t) = Z_i - \left(\frac{\sum_{j=1}^{n} Y_j(t) Z_j e^{Z_j \hat{\beta}}}{\sum_{j=1}^{n} Y_j(t) e^{Z_j \hat{\beta}}} \right)$$

Grambsch et Therneau [46] ont remarqué que pour un coefficient $\hat{\beta}_l$ estimé à partir d'un modèle de Cox, on a la relation $E(r^*_{sl}(t)) + \hat{\beta}_l \approx \beta_l(t)$, où $r^*_{sl}(t)$ sont les résidus réduits de Shoenfeld qui valent $r^*_{sl}(t) = \left(\frac{S_2(s,\hat{\beta})}{S_0(s,\hat{\beta})} - E(s,\hat{\beta})^{\otimes 2} \right)^{-1} r_s(t)$. Ainsi, en représentant graphiquement ces résidus réduits en fonction du temps, ou d'une fonction du temps $g(t)$, et en ajoutant une courbe de lissage avec des bandes de confiance à 95 %, on doit obtenir, si l'hypothèse de proportionnalité est respectée, une droite horizontale avec les bandes de confiance ne croisant pas la droite horizontale d'ordonnée correspondant à l'estimation $\hat{\beta}_l$. Afin d'obtenir une mesure plus objective pour vérifier la proportionnalité, Grambsch et Therneau [46] ont proposé un test basé sur la corrélation des résidus réduits de Schoenfeld en fonction d'une fonction du temps. On obtient une statistique qui suit asymptotiquement une loi du χ^2 à r degrés de liberté. Il est ainsi possible de tester non seulement la constance de l'effet pour chaque covariable mais également de tester globalement la constance de tous les effets. Presque tous les tests de la constance de l'effet au cours du temps sont basés sur la même statistique, seule la fonction du temps $g(t)$ changeant. Dans cette thèse, la transformation correspondant au rang d'événements a été choisie. De nombreuses autres méthodes permettant de vérifier l'hypothèse de proportionnalité de manière graphique ou analytique ont été étudiées. Le lecteur intéressé trouvera ici quelques références [21] et [24]. En cas de non respect de l'hypothèse de proportionnalité, il est nécessaire d'avoir recours à une extension du modèle de Cox. Dans ce travail, le modèle présenté au début de ce paragraphe a été utilisé.

1.2.2.2. Absence de log-linéarité

Le modèle de Cox suppose la log-linéarité des effets des covariables quantitatives, c'est-à-dire que l'augmentation d'une unité de la variable quantitative a le même effet quelle que soit la valeur de la variable choisie. Si cette hypothèse n'est pas vérifiée, il est nécessaire d'utiliser une extension du modèle de Cox, la plus courante consistant à introduire une forme fonctionnelle. Cette extension du modèle de Cox s'écrit :

$$\lambda(t) = Y(t)\lambda_0(t)e^{f(Z)\beta} \tag{1.7}$$

où f est une fonction correspondant à la forme fonctionnelle de la variable et dont le choix dépend de connaissances antérieures ou des outils diagnostiques. Pour vérifier la justesse de cette hypothèse, Grambsch et Therneau ont proposé d'utiliser les résidus de martingale. Les résidus de martingale se définissent pour chaque sujet à partir des processus de résidus de martingale qui sont la différence entre le processus de comptage $N_i(t)$ et le compensateur $\Lambda_i(t)$:

$$M_i(t) = N_i(t) - \Lambda_i(t).$$

Les résidus de martingale sont égaux aux processus de résidus de martingale à la fin du temps de suivi pour chaque sujet. Dans le cas d'un modèle de Cox dont l'effet des covariables ne dépend pas du temps, ils s'écrivent :

$$M_i = \delta_i - \Lambda_0(\tau_i)e^{Z_i\beta}$$

où δ_i est le statut final, $\Lambda_0(\tau_i)$, le risque instantané cumulé au temps d'observation τ_i du sujet i correspondant au moment où le sujet présente l'événement d'intérêt ou est censuré et Z_i, le vecteur des covariables pour le patient i. Ces résidus ont une valeur comprise entre 1 et $-\infty$ et sont donc asymétriques. Ils présentent comme propriétés :

— $E(M_i) = 0$
— $\sum_{i=1}^{n} \hat{M}_i = 0$

Therneau et Grambsch [47] ont montré qu'en représentant graphiquement les résidus de martingale obtenus avec un modèle de Cox vide, c'est-à-dire avec le vecteur des paramètres à 0 ($\hat{\beta} = 0$), en fonction de chaque variable continue, la courbe de lissage dessine la forme fonctionnelle de la variable. Cette forme fonctionnelle peut être modélisée et introduite dans le modèle de Cox étendu pour obtenir des résidus de martingale. Afin de vérifier que le choix est approprié, la courbe de lissage de ces résidus en fonction de la variable continue est représentée graphiquement. Si elle est horizontale, alors le modèle s'ajuste correctement aux données [45].

1.2.2.3. Absence de propotionnalité des risques instantanés et de log-linéarité

En cas de non-proportionnalité et d'absence de log-linéarité pour une variable continue, il est possible d'utiliser un modèle de Cox étendu ayant les caractéristiques

des deux modèles précédents et qui s'écrit :

$$\lambda(t) = Y(t)\lambda_0(t)e^{f(Z)\beta(t)}. \tag{1.8}$$

1.3. Modèles additifs

1.3.1. Modèle d'Aalen

1.3.1.1. Définition du modèle

Le modèle d'Aalen est un modèle à risques instantanés additifs qui a été proposé par Aalen en 1989 [3]. Ce modèle repose sur l'équation suivante :

$$\lambda(t) = X(t)\alpha(t) \tag{1.9}$$

où

- $X(t)$ est une matrice de taille n (représentant le nombre de sujets) par $r+1$ (soit r=nombre de covariables plus l'ordonnée à l'origine) avec pour chaque ligne i $X_i(t) = I(t < T_i)Z_i(t)$ et $Z_i(t) = (1, Z_{i1}(t), Z_{i2}(t), ..., Z_{ir}(t))$
- $\alpha(t)$ est le vecteur des fonctions non-paramétriques de régression
- T_i est le temps jusqu'à l'événement ou la censure pour le sujet i

Dans le cadre du modèle d'Aalen, les paramètres $\alpha(t)$ sont des fonctions du temps et il n'est pas fait d'hypothèses sur leurs formes. Par conséquent le modèle est non-paramétrique. Le premier élément de $\alpha(t)$, $\alpha_0(t)$, est la fonction de risque de base et les autres éléments $(\alpha_1(t), ...\alpha_r(t))$ sont les fonctions de régression, mesurant l'effet additionnel des covariables sur le risque instantané de base. Comme les fonctions de régression peuvent varier dans le temps, il est possible d'étudier l'effet des covariables sur le risque instantané au cours du temps.

La fonction de risque instantané pour un individu i s'écrit donc :

$$\lambda_i(t) = I(t \leq T_i)(\alpha_0(t) + \alpha_1(t)z_{i1} + ... + \alpha_r(t)z_{ir}). \tag{1.10}$$

1.3.1.2. Estimation des fonctions de régression du modèle

Soient n observations indépendantes et identiquement distribuées issues d'un modèle d'Aalen, $N(t) = (N_1(t), ...N_n(t))'$ les processus de comptage correspondant à chaque sujet, $M(t) = (M_1(t), ...M_n(t))'$ les martingales et $X(t)$ la matrice des covariables. Le modèle s'écrit :

$$N(t) = \Lambda(t) + M(t) = X(t)A(t) + M(t).$$

En dérivant par rapport au temps, on obtient :

$$dN(t) = X(t)dA(t) + dM(t)$$

où

$$A(t) = \int_0^t \alpha(s)ds$$

est le vecteur des fonctions de régression cumulées. On obtient un estimateur de $A(t)$ (correspondant aux moindres carrés ordinaires) :

$$\hat{A}(t) = \int_0^t \left(X(s)'X(s)\right)^{-1} X(s)'dN(s)$$

que l'on peut réécrire :

$$\hat{A}(t) = \sum_{T_k \leq t} G(T_k) I_k \tag{1.11}$$

où T_k sont les temps d'événements, $G(t)$ est une inverse généralisée de $X(t)$ et I_k est un vecteur de 0 excepté 1 quand un sujet fait un événement au temps T_k. Quand $X(t)$ n'est plus de plein rang, l'estimation de $\hat{A}(t)$ est égale à $\hat{A}(t_{-1})$ calculé au dernier temps d'événement pour lequel $X(t)$ est de plein rang. Cet estimateur du risque cumulé a pour propriété, en présence d'une seule variable qualitative explicative, d'être égal pour chacune des modalités, à l'estimateur de Nelson-Aalen pour chacun des sous-groupes. Huffer et McKeague [22] ont proposé une autre méthode d'estimation basée sur les moindres carrés pondérés. Dans ce travail, seule l'estimation basée sur les moindres carrés ordinaires a été étudiée. L'analyse des fonctions de régression se fait en représentant chaque fonction $\hat{A}(t)$ en fonction du temps, ce qui permet de visualiser l'effet propre de chaque covariable sur le risque instantané cumulé. La théorie des martingales et des intégrales stochastiques permet d'obtenir une estimation de la matrice de variance covariance de $\hat{A}(t)$ qui s'écrit :

$$\hat{\Omega}(t) = \sum_{T_k \leq t} G(T_k) diag(I_k) G(T_k)'. \tag{1.12}$$

On peut ainsi représenter les deux courbes de la bande de confiance à 95 % des fonctions de régression en calculant $\hat{A}_l(t) \pm 1{,}96\sqrt{\hat{\Omega}_{ll}(t)}$, ce qui permet de vérifier graphiquement l'absence d'effet d'une variable si l'axe des abscisses est inclus dans la bande de confiance ou la constance de l'effet de la covariable si l'on peut tracer une droite dans la bande de confiance sans qu'elle ne croise l'une des deux courbes de cette bande.

1.3.1.3. Interprétation des fonctions de régression

La fonction $A_0(t)$ représente le risque instantané cumulé de base, c'est-à-dire le risque instantané cumulé d'un patient dont toutes les covariables sont nulles. Chaque fonction $A_l(t)$ représente le risque instantané cumulé dû à l'effet de chaque covariable Z_l. En présence d'une covariable binaire d'exposition Z_1 et de deux sujets ($i = 1,2$)

avec $Z_{11} = 1$ et $Z_{21} = 0$, on a alors une différence de leurs risques instantanés cumulés :

$$\Lambda_1(t) - \Lambda_2(t) = A_1(t)$$

qui représente le risque instantané cumulé de présenter un événement dû à cette covariable. Cette fonction du risque instantané cumulé a l'avantage de ne pas dépendre du risque instantané de base et de pouvoir varier librement dans le temps. Ce risque instantané cumulé se rapproche de l'incidence due à une exposition quand ce risque instantané est faible [48]. En effet, on a l'incidence due à une exposition Inc qui vaut

$$Inc(t) = 1 - S(t) = 1 - e^{-\Lambda(t)} \approx \Lambda(t)$$

quand $\Lambda(t) < 0,1$. Le modèle d'Aalen présente toutefois deux inconvénients calculatoires. Le premier est qu'il n'impose pas une positivité du risque instantané, et qu'il se peut que dans certains cas le risque instantané cumulé soit décroissant à certains moments voire même négatif pour certains sujets. Il en résulte soit une survie croissante, soit une survie supérieure à 1, ce qui s'oppose à la définition de la fonction de survie. Toutefois, ce problème se rencontre rarement excepté quand l'effectif est trop petit. Le second problème réside dans la méthode d'estimation qui nécessite une inversion matricielle qui peut ne pas être estimable en présence d'un trop grand nombre de covariables ou à la fin du suivi.

1.3.1.4. Test des fonctions de régression

Aalen [3] a proposé un test des paramètres du modèle, c'est à dire un test de l'hypothèse selon laquelle une fonction $\alpha_l(t)$ est égale à 0. On a donc : $H_{0l} : \alpha_l(t) = 0 \ \ \forall t$. Une statistique pour tester l'hypothèse H_{0l} est donnée par l'élément l de U :

$$U = \sum_{T_k \leq t} K(T_k) G(T_k) I_k$$

où $K(t)$ est une matrice diagonale représentant des poids. Aalen propose deux possibilités pour $K(t)$:

— soit chaque élément de la diagonale représente le nombre N de sujets encore à risque au temps t. On peut alors substituer $K(t)$ par le scalaire $N(t)$.
— soit les éléments de la matrice $(X'X)^{-1}$ sont pondérés par un poids inverse à celui de la variance : $K(t) = \left\{ diag\left((X(t)'X(t)) \right)^{-1} \right\}^{-1}$

C'est cette seconde possibilité qui est préconisée par Aalen et qui est utilisée dans ce travail. Un estimateur de la matrice de variance covariance de U est :

$$V = \sum_{T_k \leq t} K(T_k) G(T_k) diag(I_k) G(T_k)' K(T_k).$$

Pour tester une hypothèse H_{0l}, on peut utiliser la statistique $U_l V_{ll}^{-1/2}$ qui suit une loi normale. Pour tester simultanément plusieurs hypothèses H_{0l} avec $l \in R = 1, 2, ..., r$,

on peut utiliser la statistique $U_R' V_R^{-1} U_R$ qui suit une loi du χ^2 à s (avec $s \leq r$) degrés de liberté correpondant au nombre de fonctions testées.

1.3.2. Extension du modèle d'Aalen

Le modèle d'Aalen est non-paramétrique et permet ainsi de modéliser l'effet d'une variable au cours du temps. Il n'y a donc pas d'hypothèses à faire sur la constance de l'effet au cours du temps. Toutefois, le modèle d'Aalen suppose la linéarité des effets des covariables quantitatives, c'est-à-dire que l'augmentation d'une unité de la variable quantitative a le même effet quelle que soit la valeur de la variable choisie. Si cette hypothèse n'est pas vérifiée, il est nécessaire d'utiliser une extension du modèle d'Aalen permettant introduire une forme fonctionnelle. Cette extension du modèle d'Aalen s'écrit pour une variable quantitative Z :

$$\lambda(t) = I(t \leq T_i)\left(\beta_0(t) + \beta_1(t) f(Z)\right) \tag{1.13}$$

où f est une fonction correspondant à la forme fonctionnelle de la variable et dont le choix dépend de connaissances antérieures ou des outils diagnostiques qui seront étudiés au chapitre 2.

1.3.3. Autres modèles additifs

1.3.3.1. Modèle semi-paramétrique de Lin

Lin et Ying [26] ont proposé en 1994 un modèle à risques instantanés additifs analogue au modèle de Cox dans un cadre additif reposant sur l'équation :

$$\lambda(t) = Y(t)\left(\lambda_0(t) + Z\gamma\right) \tag{1.14}$$

où

- $\lambda_0(t)$ est la fonction de risque instantané de base. Elle correspond au risque instantané des sujets ayant leurs covariables fixées à la valeur de référence. C'est une fonction qui est non-paramétrique et qui dépend uniquement du temps.
- les paramètres γ sont constants au cours du temps.

Comme pour le modèle de Cox, l'estimation des paramètres est basée sur la maximisation de la vraisemblance partielle et la matrice de variance covariance est obtenue avec la Hessienne de cette vraisemblance partielle. Ce modèle, tout comme le modèle d'Aalen, suppose la linéarité des effets des covariables quantitatives.

1.3.3.2. Modèle semi-paramétrique de McKeague et Sasieni

S'inspirant des travaux de Lin et Ying, McKeague et Sasieni [33] ont proposé la même année un modèle à risques instantanés additifs permettant à certaines covariables d'avoir un effet variant au cours du temps et aux autres d'avoir un effet constant. Il

s'écrit :

$$\lambda(t) = Y(t)\left(Z_1\alpha(t) + Z_2\gamma\right) \tag{1.15}$$

où

- Z_1 correspond à la matrice des covariables incluant l'ordonnée à l'origine dont les effets varient au cours du temps
- Z_2 correspond à la matrice des covariables dont les effets sont constants au cours du temps
- $\alpha(t)$ est le vecteur des fonctions de régression non-paramétriques
- les paramètres γ sont constants au cours du temps.

L'estimation des paramètres γ se fait par maximisation de la vraisemblance et l'estimation des fonctions de régression non-paramétriques $\beta(t)$ par la méthode des moindres carrés.

1.4. Modèles multiplicatifs et additifs

Afin de prendre en compte simultanément des covariables ayant un effet multiplicatif et d'autres ayant un effet additif, différents modèles ont été développés.

1.4.1. Modèle de Cox-Aalen

Scheike et Zhang [43] ont proposé en 2002 de modéliser le risque instantané comme un modèle de Cox en remplaçant le risque instantané de base par un modèle d'Aalen. Il s'écrit :

$$\lambda(t) = Y(t)\left(Z_1\alpha(t)\right)e^{Z_2\beta}$$

où

- Z_1 correspond à la matrice des covariables incluant l'ordonnée à l'origine dont les effets additifs varient au cours du temps
- Z_2 correspond à la matrice des covariables dont les effets multiplicatifs sont constants au cours du temps
- $\alpha(t)$ est le vecteur des fonctions de régression non-paramétriques
- les paramètres β sont constants au cours du temps.

C'est une extension du modèle de Cox stratifié. L'estimation des paramètres se fait en maximisant la vraisemblance partielle et l'estimation des fonctions de régression à partir de la méthode des moindres carrés ordinaires.

1.4.2. Autres modèles multiplicatifs et additifs

Lin et Ying [27] ont étudié en 1995 un modèle ayant un risque instantané qui s'écrit :

$$\lambda(t) = Y(t)\left(Z_1\gamma + \lambda_0 e^{Z_2\beta}\right).$$

Ce modèle ne permet pas de prendre en compte des variables dont l'effet dépend du temps. Martinussen et Scheike [31] ont développé ce modèle en permettant à la partie additive de modéliser les covariables dont les effets dépendent du temps

$$\lambda(t) = Y(t)\left(Z_1\alpha(t) + \lambda_0 e^{Z_2\beta}\right).$$

L'estimation des paramètres et des fonctions de régression se fait à partir du score. Malheureusement, ces modèles ne sont pas implémentés dans les logiciels de statistiques et ils ne sont donc pas très utilisés.

2. Modélisation du risque instantané

Sommaire

2.1 Outils diagnostiques . . . 30
2.1.1 Pseudo-observations . . . 30
2.1.2 Graphiques d'Arjas . . . 32
2.1.3 Processus de résidus de martingale . . . 33
2.2 Proposition d'une stratégie de modélisation des modèles à risques instantanés multiplicatifs . . . 34
2.3 Proposition d'une stratégie de modélisation des modèles à risques instantanés additifs . . . 35
2.4 Résultats . . . 35
2.4.1 Modélisation multiplicative . . . 36
2.4.1.1 Hypothèse de proportionnalité des risques . . . 37
2.4.1.2 Qualité de l'ajustement . . . 38
2.4.1.3 Ajustement du modèle à risques instantanés multiplicatifs multivarié . . . 40
2.4.2 Modélisation additive . . . 45
2.4.2.1 Hypothèse de linéarité . . . 45
2.4.2.2 Hypothèse de la constance de l'effet . . . 47
2.4.2.3 Qualité de l'ajustement . . . 49
2.5 Discussion . . . 53

2.1. Outils diagnostiques

2.1.1. Pseudo-observations

Les outils diagnostiques concernant le modèle de Cox présentés au premier chapitre supposent que l'une des deux hypothèses est respectée (proportionnalité des risques instantanés ou log-linéarité) pour vérifier la seconde. En effet, pour la log-linéarité, les résidus de martingale sont calculés pour chaque sujet à son temps d'événement ou de censure mais sont représentés sur le même graphique afin d'obtenir une courbe de lissage. Afin de vérifier simultanément les deux hypothèses, il existe différentes approches. Sasieni et Winnett [42] ont proposé une nouvelle sorte de résidus basés sur les résidus de martingale et de Schoenfeld permettant de vérifier les deux hypothèses simultanément. Abrahamowicz et McKenzie [5] ont quant à eux proposé un modèle

permettant de prendre en compte directement les variables non-log-linéaires dont l'effet dépend du temps avec des fonctions spline.

Une autre approche a été proposée par Andersen [7] basée sur les travaux de Quenouille [41] portant originellement sur les techniques de réduction des biais afin de réaliser des modèles linéaires généralisés avec des données de survie. En effet, dans le cadre de l'analyse de survie, du fait de la présence de censures, on n'observe pas pour chaque sujet la variable d'intérêt T^*, à savoir le délai avant la survenue de l'événement d'intérêt. En absence de données censurées, on pourrait utiliser un modèle de régression pour une fonction de la variable d'intérêt $f(T^*)$ et vérifier les hypothèses avec les méthodes graphiques utilisées habituellement. En présence de données censurées, il est possible de remplacer la fonction de la variable d'intérêt par une pseudo-observation. En effet, si la moyenne du paramètre d'intérêt est $\theta = E(f(T^*))$ que l'on estime par $\hat{\theta}$, alors, on définit la pseudo-observation pour la variable $f(T^*)$ comme étant la différence entre n fois le paramètre calculé sur l'ensemble des données et $n-1$ fois le paramètre calculé sur l'ensemble des données excepté un sujet. Elle s'écrit : $\hat{\theta}_i(X) = n\hat{\theta}(X) - (n-1)\hat{\theta}^{-i}(X)$ où

— $\hat{\theta}$ est estimé avec tout l'échantillon

— $\hat{\theta}^{-i}$ est estimé avec tout l'échantillon moins l'observation i.

Dans le cadre de l'analyse de survie, en prenant $\hat{\theta} = \hat{S}(t)$ qui est habituellement une estimation de Kaplan-Meier, on a $\hat{S}_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{-i}(X)$. On obtient donc une pseudo-observation par sujet et par temps. Les pseudo-observations ont pour propriété : $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} S_i(t)$. Les pseudo-observations se calculent aussi sur d'autres estimateurs comme le temps de survie moyen restreint [8] et l'incidence cumulée dans le cadre des risques concurrents [25]. Pour obtenir un estimateur directement utilisable avec les modèles à risques instantanés, nous avons calculé ces pseudo-observations avec l'estimateur de Nelson-Aalen. Les résultats obtenus montrent qu'elles semblent présenter un biais. En effet, bien qu'elles convervent la propriété observée avec l'estimateur de Kaplan-Meier, elles ont l'inconvénient de prendre de grandes valeurs au cours du temps, l'estimateur de Nelson-Aalen n'étant pas, à la différence de celui de Kaplan-Meier, borné.

Afin d'étudier la variation de la survie en fonction d'une variable continue Z, les pseudo-observations vont être utilisées à la place de la survie et être représentées graphiquement par un nuage de points en fonction d'une covariable. Un lissage permet ainsi de visualiser l'effet de la covariable sur la survie. Pour vérifier les hypothèses du modèle de Cox (risques instantanés proportionnels et log-linéarité), il est nécessaire de procéder à une transformation permettant d'exprimer directement l'effet des covariables sur la survie. En effet, pour le modèle de Cox, on a $S(t|Z) = e^{-\Lambda_0(t)e^{\beta Z}}$ donc $ln(-ln(S(t|Z)) = ln(\Lambda_0(t)) + \beta Z$. Afin de visualiser l'évolution de l'effet dans le temps, il est possible de représenter les pseudo-observations en fonction de la covariable Z et du temps en trois dimensions et d'ajouter une surface de lissage. Toutefois, les graphiques en trois dimensions étant difficiles à interpréter, il est d'usage de représenter ces courbes de lissage à différents temps, comme par exemple les neuf déciles des temps d'événements. Ces graphiques permettent de vérifier les deux hypothèses :

— Si les deux hypothèses sont vérifiées, les courbes de lissage sont des droites parallèles dont la pente représente la valeur du paramètre β et l'ordonnée à l'origine vaut $ln(\Lambda_0(t))$ à chaque temps.
— Si seule l'hypothèse de proportionnalité n'est pas respectée, les courbes de lissage sont des droites non-parallèles, leur pente représentant la valeur du paramètre β aux différents temps.
— Si seule l'hypothèse de log-linéarité n'est pas respectée, les courbes de lissage obtenues sont des courbes parallèles indiquant la forme fonctionnelle de la variable.
— Si les deux hypothèses ne sont pas respectées, les courbes de lissage obtenues sont des courbes non-parallèles indiquant la forme fonctionnelle de la variable à chaque temps.

L'avantage de cette approche est qu'elle est basée sur une estimation non-paramétrique de la survie et qu'elle peut donc être utilisée également pour les modèles à risques instantanés additifs. Pour cela, il est nécessaire d'utiliser un modèle de Lin qui fait des hypothèses équivalentes au modèle de Cox, à savoir la constance de l'effet au cours du temps et la linéarité de l'effet. Pour le modèle de Lin, on peut également exprimer l'effet des covariables sur la survie en utilisant une transformation. En effet, celle-ci vaut $S(t|Z) = e^{-\Lambda_0(t)-\gamma Z t}$ et on a donc $\frac{-ln(S(t|Z))}{t} = \frac{-\Lambda_0(t)(t)}{t} + \gamma Z$. La représentation graphique de ces pseudo-observations permet de vérifier les deux hypothèses :

— En cas de constance de l'effet et de linéarité de la variable, les courbes de lissage sont des droites parallèles, leur pente représentant la valeur du paramètre γ et l'ordonnée à l'origine aux différents temps valant $\frac{-\Lambda_0(t)(t)}{t}$.
— Si seule l'hypothèse de la constance de l'effet n'est pas respectée, les courbes de lissage sont des droites non-parallèles, leur pente représentant la valeur du paramètre γ aux différents temps.
— Si seule l'hypothèse de linéarité n'est pas respectée, les courbes de lissage obtenues sont des courbes parallèles indiquant la forme fonctionnelle de la covariable.
— Si les deux hypothèses ne sont pas respectées, les courbes de lissage obtenues sont des courbes non-parallèles indiquant la forme fonctionnelle de la covariable à chaque temps.

2.1.2. Graphiques d'Arjas

Afin d'étudier l'ajustement du modèle de Cox aux données, Arjas [9] a proposé de représenter graphiquement le nombre d'événements estimés par le modèle en fonction du nombre d'événements observés. Si le modèle s'ajuste correctement aux données, les deux valeurs doivent être proches et la courbe obtenue doit se superposer à la diagonale. Pour vérifier l'ajustement en fonction d'une covariable, on représente une courbe pour chacune des modalités pour les variables qualitatives ou pour des groupes de patients constitués en divisant les variables quantitatives en strates. Habituellement quatre strates constituées par les quartiles de la variable quantitative sont

représentées. Si cette approche fut développée à l'origine pour vérifier l'ajustement du modèle de Cox aux données, elle peut être également utilisée pour tous les autres types de modèles de régression à risques instantanés. Aussi, Aalen a proposé d'utiliser cette méthode pour vérifier l'ajustement de son modèle aux données. Toutefois, s'il est possible de vérifier l'ajustement du modèle aux données pour chacune des strates constituées en divisant les variables quantitatives, cette approche ne permet pas de vérifier l'ajustement du modèle aux données pour les variables qualitatives car l'estimation du nombre d'événements par modalités des variables qualitatives égale le nombre d'événements observés. En effet, on a vu que l'estimateur de Nelson-Aalen a pour propriété $\sum_{i=1}^{n} \hat{H}(t_i) = \sum_{i=1}^{n} N(t_i)$ et qu'il est égal à l'estimateur du risque cumulé du modèle d'Aalen $\hat{A}(t)$ en présence d'une seule variable qualitative explicative pour chacune des modalités. Il est en revanche possible et utile d'utiliser des graphiques d'Arjas pour vérifier l'adéquation du modèle de Lin pour les variables quantitatives et qualitatives.

2.1.3. Processus de résidus de martingale

Si les graphiques d'Arjas sont des outils utiles pour vérifier l'ajustement du modèle d'Aalen au données, ils ne permettent pas de suivre l'évolution du nombre d'événements estimés et observés au cours du temps. Pour cela, les processus de résidus de martingale $M_i(t)$, originellement proposés dans le cadre du modèle de Cox, peuvent également être calculés avec un modèle d'Aalen. Comme on l'a vu au chapitre précédent, ils valent la différence entre le processus de comptage $N_i(t)$ et le compensateur $\Lambda_i(t)$. Ces $M_i(t)$ sont des martingales exactes et ont pour propriété d'avoir à chaque temps pour l'ensemble des sujets une somme nulle :$\sum_{i=1}^{n} M_i(t) = \sum_{i=1}^{n} N_i(t) - \sum_{i=1}^{n} \Lambda_i(t) = 0$. Le vecteur des processus de résidus de martingale se calcule dans le cadre d'un modèle d'Aalen avec l'équation suivante :

$$M(t) = \sum_{T_k \leq t} (J - X(T_k)G(T_k))\, I_k.$$

J est la matrice identité de taille $n \times n$. Afin de vérifier l'ajustement du modèle aux données, il convient de diviser l'échantillon en différents groupes, habituellement en quatre groupes basés sur les quartiles d'une variable quantitative, et de représenter graphiquement ces processus de résidus de martingale en fonction du temps. En cas de bon ajustement, les courbes doivent rester proches de l'axe des abcisses et leurs bandes de confiance ne doivent pas croiser cet axe. Pour calculer ces processus de résidus de martingale, on utilise l'estimateur suivant :

$$M^K(t) = KM(t)$$

avec K une matrice de taille $k \times n$ composée de k lignes indicatrices indiquant pour chaque sujet son appartenance au groupe k avec un 1, 0 sinon, et pour calculer leur

variance l'estimateur qui s'écrit :

$$Var^K(t) = \sum_{T_k \leq t} K\left(J - X(T_k)G(T_k)\right) I_k^D \left(J - X(T_k)G(T_k)\right)' K'$$

où I_k^D est la matrice de taille $n \times n$ dont la diagonale est égale à I_k. Il est également possible de réaliser un test pour savoir si l'ensemble des k processus de résidus de martingale s'éloignent significativement de l'axe des abscisses car la quantité $\left(M^{K-1}(t)\right)' \left(Var^{K-1}(t)\right)^{-1} \left(M^{K-1}(t)\right)$ suit asymptotiquement une loi du χ^2 à $k-1$ degrés de liberté. Ce test peut également être réalisé pour chacune des modalités k.

Comme ces processus ne sont pas disponibles avec le logiciel R et ses progiciels, il fut nécessaire d'implémenter cette approche en langage R.

2.2. Proposition d'une stratégie de modélisation des modèles à risques instantanés multiplicatifs

Afin d'ajuster correctement un modèle à risques instantanés multiplicatifs, il est proposé d'utiliser la stratégie suivante :

1. Vérifier l'hypothèse de log-linearité pour chacune des covariables continues grâce à la représentation graphique de leurs résidus de martingale calculés avec un modèle de Cox vide. En cas d'absence de log-linéarité, la forme de la courbe de lissage permet d'aider à modéliser la forme fonctionnelle de la covariable. La représentation graphique des résidus de martingale calculés avec le modèle de Cox étendu avec la modélisation de la forme fonctionnelle permet de vérifier que celle-ci est adéquate. Si ce n'est pas le cas, il faut changer de forme fonctionnelle jusqu'à en trouver une qui satisfasse la log-linéarité.
2. Vérifier l'hypothèse de la proportionnalité des risques instantanés pour chaque covariable en utilisant les résidus de Schoenfeld. En cas de non-proportionalité, un paramètre dépendant du temps doit être modélisé. L'hypothèse de proportionalité doit être vérifiée pour chaque paramètre de la fonction retenue. Si ce n'est pas le cas, il faut changer de fonction du temps jusqu'à en trouver une qui satisfasse la log-linéarité.
3. Vérifier simultanément les hypothèses de proportionnalité des risques instantanés et de log-linearité en représentant graphiquement le logarithme de l'opposé du logarithme des pseudo-observations en fonction de la covariable continue. Si les deux hypothèses ne sont pas respectées simultanément, il faut changer de fonction du temps ou de forme fonctionnelle jusqu'à en trouver qui satisfasse les deux hypothèses.
4. Vérifier l'ajustement de chaque covariable grâce à la repésentation graphique d'Arjas. Répéter ces étapes jusqu'à ce que les hypothèses de proportionnalité des risques instantanés et de log-linearité soient respectées avec des modèles univariés.

5. Vérifier l'hypothèse de la proportionnalité du modèle ajusté avec toutes les covariables grâce au test de la corrélation entre les résidus de Schoenfeld et le rang des temps d'événement. Si l'hypothèse de proportionnalité n'est pas respectée pour toutes les covariables, il est nécessaire de recommencer cette étape en changeant l'effet.

2.3. Proposition d'une stratégie de modélisation des modèles à risques instantanés additifs

Afin d'ajuster correctement un modèle à risques instantanés additifs, il est proposé d'utiliser la stratégie suivante :

1. Choisir la forme fonctionnelle pour chaque covariable continue en utilisant la représentation graphique de la courbe de lissage de l'opposé du logarithme des pseudo-observations divisé par le temps en fonction de la covariable continue. Ensuite, la vérification de l'ajustement du modèle avec la forme fonctionnelle retenue se fait en représentant les processus de résidus de martingale au cours du temps pour chaque groupe obtenu en divisant la variable continue en quatre strates et avec des tests du χ^2 comparant le nombre d'événements observés par groupe avec le nombre d'événements estimés par le modèle et par groupe. Si les tests sont statistiquement significatifs, une autre forme fonctionnelle doit être choisie et testée jusqu'à en trouver une qui satisfasse la linéarité.
2. Vérifier la constance de l'effet de chaque covariable en représentant graphiquement les risques instantanés cumulés estimés avec le modèle de Lin et le modèle d'Aalen. Si les effets de toutes les covariables sont constants, il est possible d'utiliser le modèle de Lin, sinon, il faut utiliser le modèle d'Aalen.
3. Vérifier l'ajustement du modèle aux données pour chaque covariable continue en utilisant les graphiques d'Arjas. En cas de mauvais ajustement, changer de forme fonctionnelle jusqu'à l'obtention d'une forme fonctionnelle qui satisfasse l'ajustement correct.
4. Vérifier l'hypothèse de linéarité du modèle ajusté avec toutes les covariables en utilisant les graphiques et les tests des processus de résidus de martingale pour toutes les covariables continues. Si les tests sont statistiquement significatifs, une autre forme fonctionnelle doit être choisie et testée jusqu'à en trouver une qui satisfasse la linéarité.

2.4. Résultats

Cette partie présente l'utilisation de la stratégie proposée dans les deux sections précédentes pour analyser les données de survie issues de la base de données TRACE incluse dans le progiciel `timereg`. Cette base de données contient 1878 patients et

est issue d'une étude comportant 4259 patients hospitalisés pour un infarctus du myocarde à Copehangue au Danemark entre 1977 et 1988 et qui furent suivis jusqu'à leur décès ou leur censure. Les covariables qui ont été étudiées sont l'âge, qui est une variable continue codée `age`, la présence d'une insuffisance cardiaque, qui est une variable binaire codée `ic`, le sexe, qui est une variable binaire codée `sexe`, le diabète, qui est une covariable binaire codée `dia` et la fibrillation ventriculaire, qui est une covariable binaire codée `fv`. L'estimation des fonctions de régression du modèle d'Aalen nécessitant l'absence d'ex-aequo quant à la durée de suivi, un nombre aléatoire a été ajouté à tous ces temps de suivi. L'âge moyen des patients était de 67,0 ans avec un écart-type de 11,4 ans, 52,29 % des patients avaient une insuffisance cardiaque, 69,54 % étaient des femmes, 10,01 % étaient diabétiques et 7,24 % ont eu une fibrillation ventriculaire. La survie médiane était de 6,52 ans ($IC_{95\%} = [6,09; 7,25]$). Cette base de données a été utilisée car elle présente des variables continues et binaires et certaines, notamment la fibrillation ventriculaire, dont on sait que son effet n'est pas constant au cours de temps.

2.4.1. Modélisation multiplicative

Hypothèse de log-linéarité

Afin de vérifier l'hypothèse de log-linéarité pour l'âge, les résidus de martingales issus d'un modèle de Cox vide ont été calculés et représentés sur la figure 2.1 en fonction de l'âge. La courbe de lissage montre que l'effet est globalement exponentiel. Elle correspond également à une fonction quadratique.

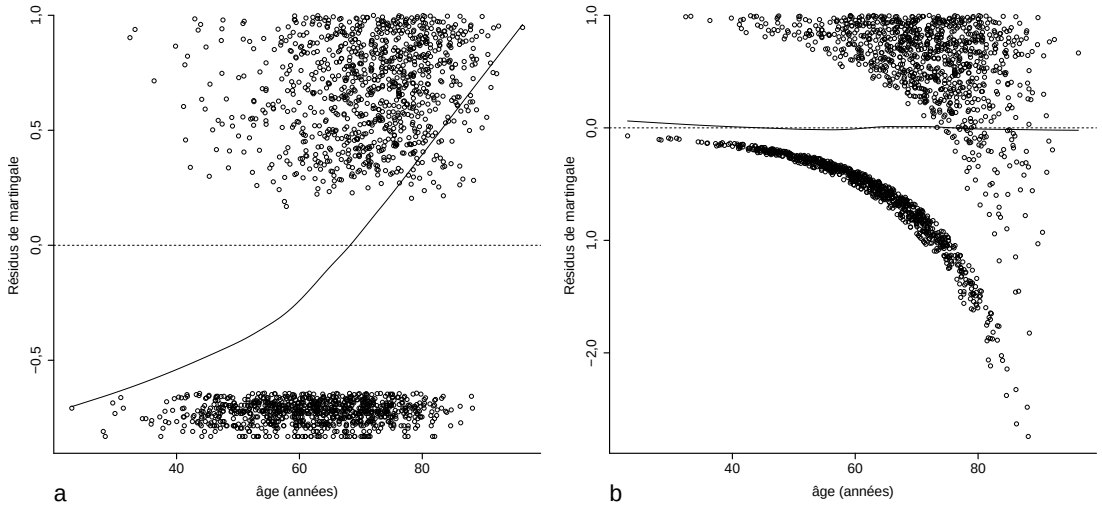


FIGURE 2.1. – Résidus de martingale en fonction de l'âge modélisé avec (a) un modèle de Cox vide et avec (b) un modèle de Cox ajusté avec l'exponentielle de l'âge sur 100. Une courbe de lissage est ajoutée en trait plein sur chaque graphique.

L'âge a été divisé par 100 afin d'obtenir des valeurs de son exponentielle qui ne soient pas trop grandes et un modèle de Cox a été ajusté avec l'exponentielle de l'âge. La courbe de lissage des résidus de martingales obtenus avec ce modèle en fonction de l'âge était une droite horizontale et par conséquent l'exponentielle était une forme fonctionnelle adéquate. Comme la courbe de lissage obtenue avec le modèle de Cox vide semblait également proche d'une fonction quadratique, un modèle de Cox avec une fonction quadratique a été réalisé et la courbe de lissage des résidus de martingales obtenus avec ce modèle en fonction de l'âge était très similaire à celle obtenue avec l'exponentielle (figure .1). Il n'est pas possible de distinguer visuellement la courbe qui est la plus proche de l'horizontale. Par conséquent, un calcul de l'AIC réalisé avec chacun des deux modèles de Cox montrait que le modèle ajusté avec un effet quadratique avai un AIC légèrement supérieur (13 533,73) à celui du modèle ajusté avec l'exponentielle de l'âge (13 531,83). C'est donc cette dernière forme qui fut retenue.

2.4.1.1. Hypothèse de proportionnalité des risques

Les tests de corrélation entre les résidus de Schoenfeld et les rangs des temps d'événement (tableau 2.1) montrait que l'hypothèse de proportionnalité des risques instantanés n'était pas rejetée pour les effets de l'exponentielle de l'âge divisé par 100, du sexe et du diabète mais qu'elle l'était pour les effets de l'insuffisance cardiaque et de la fibrillation ventriculaire.

covariable	degré de significativité
exp(age/100)	0,29
sexe	0,66
ic	$1,8 \times 10^{-4}$
dia	0,23
fv	$9,5 \times 10^{-10}$

Tableau 2.1. – Significativité des tests de corrélation entre les résidus de Schoenfeld et les rangs des temps d'événement.

La représentation graphique de ces corrélations a permis d'aider à trouver une fonction du temps pour ces deux covariables. En effet, celle-ci (figure 2.2) montre que l'effet de l'insuffisance cardiaque était linéairement décroissant et pouvait donc être modélisé avec une fonction linéaire du temps. Le modèle de Cox étendu suivant fut proposé : $\lambda(t|ic) = \lambda_0(t)e^{ic\times(\beta_{ic}+\beta_{ict}\times t)}$ et le test de corrélation des résidus de Schoenfeld ne rejetait pas l'hypothèse nulle pour les deux paramètres β_{ic} et β_{ict} avec un degré de significativité de 0,85 et 0,41 respectivement. Cette fonction du temps fut donc retenue. Concernant la fibrillation ventriculaire, l'effet décroissait fortement jusqu'à 0,64 an environ puis restait stable et nul pour le reste du suivi.

Une fonction linéairement décroissante puis constante avec une rupture de pente semblait permettre de modéliser cet effet. Le modèle de Cox étendu suivant fut proposé : $\lambda(t|fv) = \lambda_0(t)e^{fv\times(\beta_{fv}+\beta_{fvt}\times t+\beta_{fvt}\times(t-n)I(t>n)}$ où n est le nœud correpondant au temps de la rupture de pente. Comme il est difficile de repérer graphiquement le meilleur temps de la rupture de la pente, une méthode basée sur l'AIC en faisant varier n de 0 à 2,4 par pas de 0,01 fut utilisée et l'on obtint une valeur minimale d'AIC de 13 824,78 pour $n = 0,15$. Le test de corrélation des résidus de Schoenfeld ne rejetait pas l'hypothèse nulle pour les trois paramètres β_{fv}, β_{fvt} et β_{fvt2} avec un degré de significativité de 0,60, 0,66 et 0,66 respectivement. C'est donc cette fonction avec cette valeur de nœud qui fut retenue.

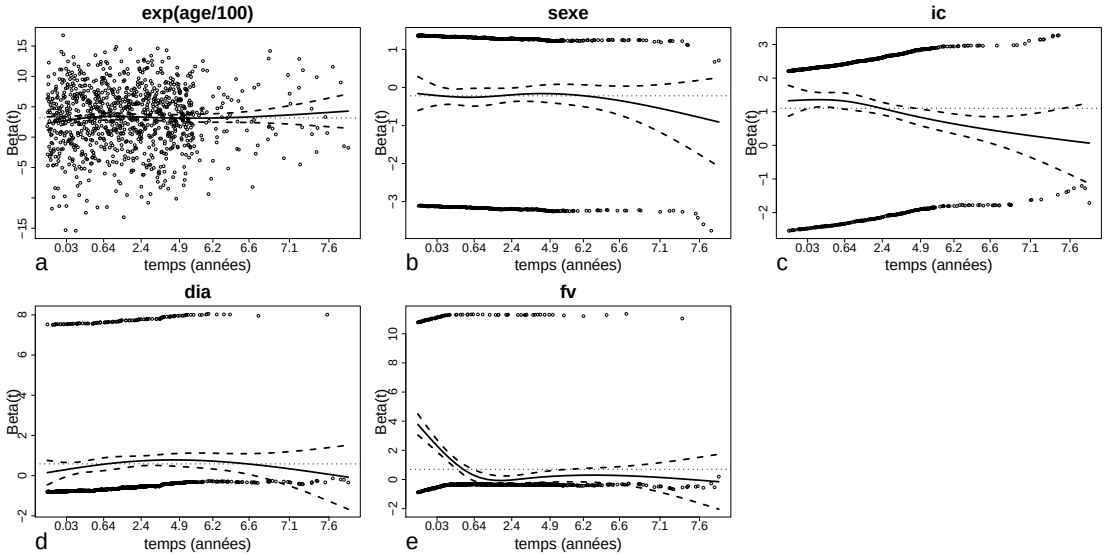


FIGURE 2.2. – Résidus de Schoenfeld en fonction du rang des événements pour (a) l'exponentielle de l'âge/100, (b) le sexe, (c) l'insuffisance cardiaque, (d) le diabète, et (e) la fibrillation ventriculaire. Une courbe de lissage en trait plein correspondant à l'estimation du paramètre β avec sa bande de confiance à 95 % en traits discontinus est ajoutée à chaque graphique.

2.4.1.2. Qualité de l'ajustement

L'estimation de la qualité de l'ajustement se fait à l'aide des pseudo-observations et des graphiques d'Arjas. La figure 2.3 représente graphiquement les courbes de lissage obtenues à partir des pseudo-observations tranformées par la fonction retenue à la section sur les pseudo-observations en fonction de l'âge et calculées aux neuf déciles des temps d'événement. Les courbes obtenues ne sont pas droites ce qui indique et confirme les résultats obtenus avec les résidus de martingale, à savoir que l'effet de l'âge n'était pas linéaire. De plus, elles ne sont pas totalement parallèles notamment

avant l'âge de 40 ans. Toutefois, comme les cinq courbes représentant les deuxième au sixième déciles sont globalement horizontales jusqu'à 40 ans avant de croître, cela indique une absence d'effet de l'âge les vingt premières années avant une croissance exponentielle. La forme particulière de la courbe correspondant au premier décile est sans doute due à l'absence d'événements précoces parmi les patients les plus jeunes et donc des estimations des pseudo-observations proches de 1 qui une fois transformées tendent vers $-\infty$. Cette courbe représente donc plus un artéfact qu'un effet véritablement différent des autres courbes. Par conséquent, l'effet exponentiel de l'âge divisé par 100 fut retenu.

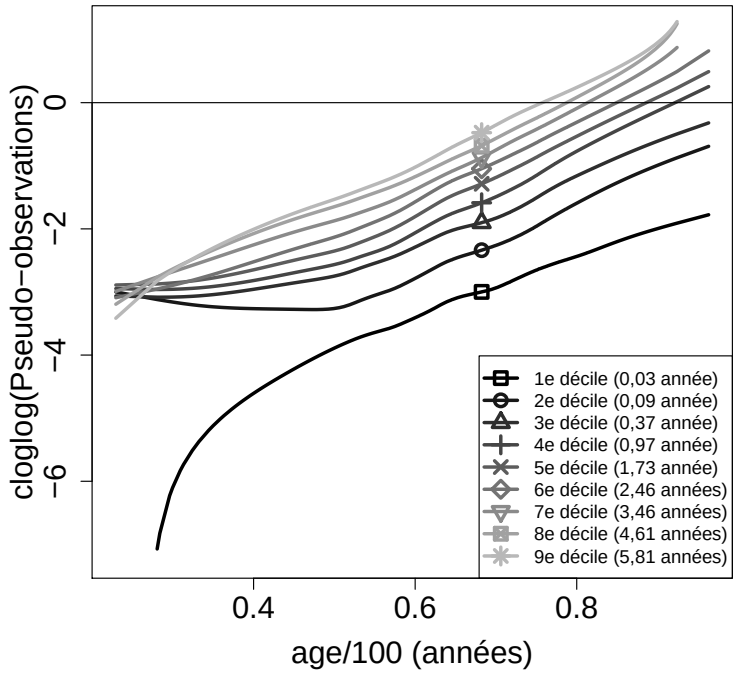


FIGURE 2.3. – Courbes de lissage obtenues à partir des pseudo-observations transformées en fonction de l'âge divisé par 100 aux neuf déciles des temps d'événement (en années).

Les graphiques d'Arjas obtenus (figure 2.4) après la division de l'âge en quatre strates définies sur ses quartiles montrent que la modélisation de l'âge de manière linéaire ne s'ajustait pas totalement correctement, notamment pour les patients des strates $]68,2;75,4]$ et $> 75,4$ car leurs courbes s'éloignent de la diagonale. La modélisation de l'âge par son exponentielle après divison par 100 présentait un meilleur ajustement puisque les quatre courbes correspondant aux quatre strates sont très proches de la diagonale confirmant ainsi les résultats obtenus avec les pseudo-observations.

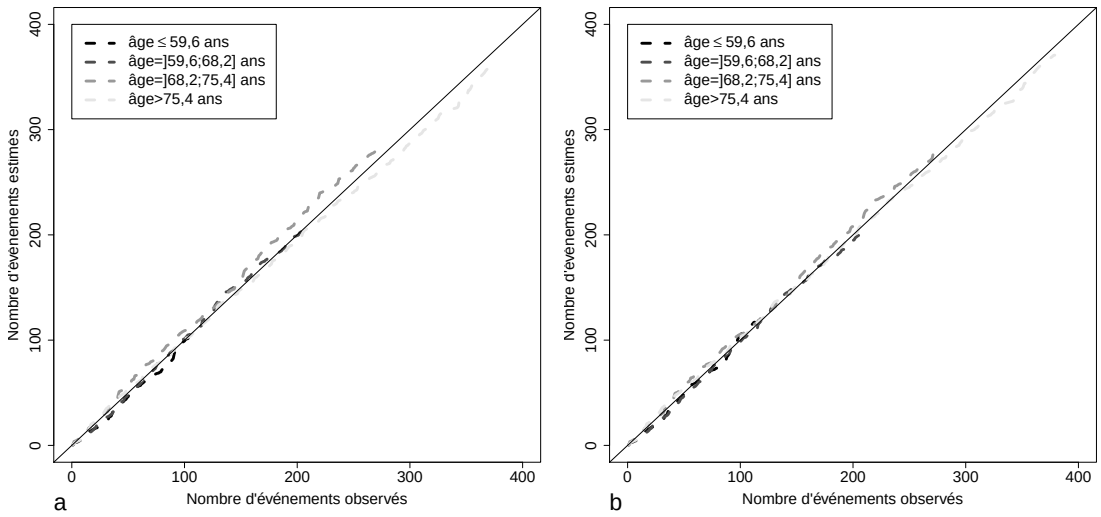


FIGURE 2.4. – Graphiques d'Arjas obtenus avec un modèle de Cox ajusté avec (a) un effet linéaire de l'âge et (b) un effet exponentiel de l'âge/100.

Concernant l'insuffisance cardiaque, quand cette covariable était modélisée sans effet dépendant du temps, le nombre d'événements estimés pour les patients sans insuffisance cardiaque était légèrement supérieur au nombre d'événements observés (figure 2.5). On observe l'inverse pour les patients avec insuffisance cardiaque. La modélisation de cette covariable avec un effet dépendant du temps permit d'obtenir des courbes qui sont pratiquement confondues avec la diagonale confirmant ainsi la bonne adéquation de cette modélisation aux données.

Les graphiques d'Arjas (figure 2.6) étudiant l'ajustement de la modélisation de la présence d'une fibrillation ventriculaire par un modèle de Cox montrent qu'en l'absence d'effet dépendant du temps, le nombre d'événements observés était assez fortement sous-estimé, mais qu'en le prenant en compte avec le modèle retenu précédemment, le nombre d'événements observés était égal au nombre d'événements estimés.

2.4.1.3. Ajustement du modèle à risques instantanés multiplicatifs multivarié

Une fois que l'hypothèse de proportionnalité a été vérifiée pour chaque covariable avec un modèle à risques instantanés multiplicatifs en analyse univariée, elle devait également être vérifiée pour l'ensemble des covariables avec le modèle à risques instantanés multiplicatifs en analyse multivariée. Le test de corrélation des résidus de Schoenfeld était statistiquement non-significatif pour l'ensemble des covariables à l'exception du diabète (tableau 2.2). La représentation graphique de la corrélation entre les résidus de Schoenfeld et le rang des temps d'événements (figure 2.7) montre qu'un effet dépendant du temps du diabète était nécessaire. Cet effet linéaire dépen-

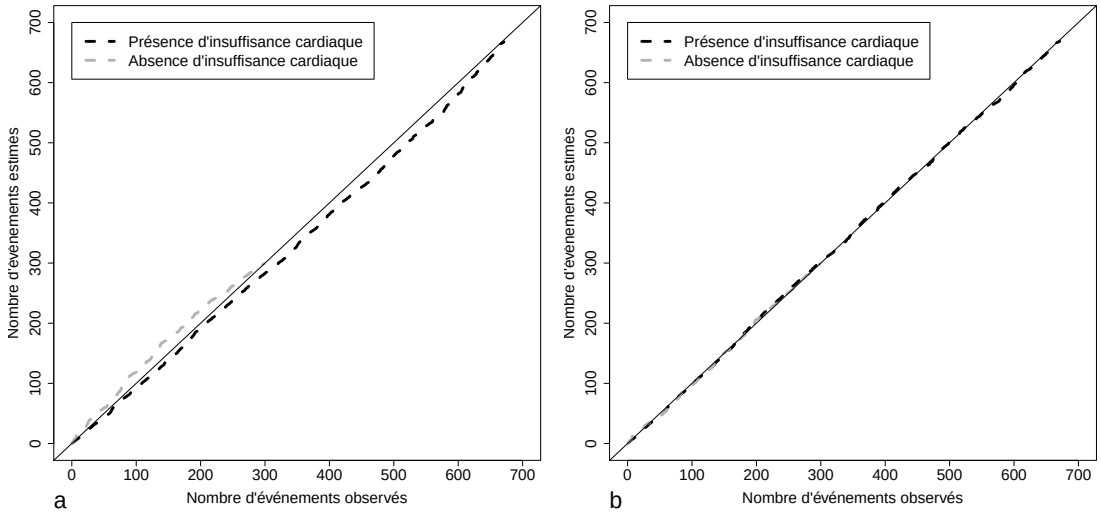


FIGURE 2.5. – Graphiques d'Arjas obtenus avec un modèle de Cox ajusté avec (a) l'insuffisance cardiaque sans effet et (b) l'insuffisance cardiaque avec un effet dépendant du temps.

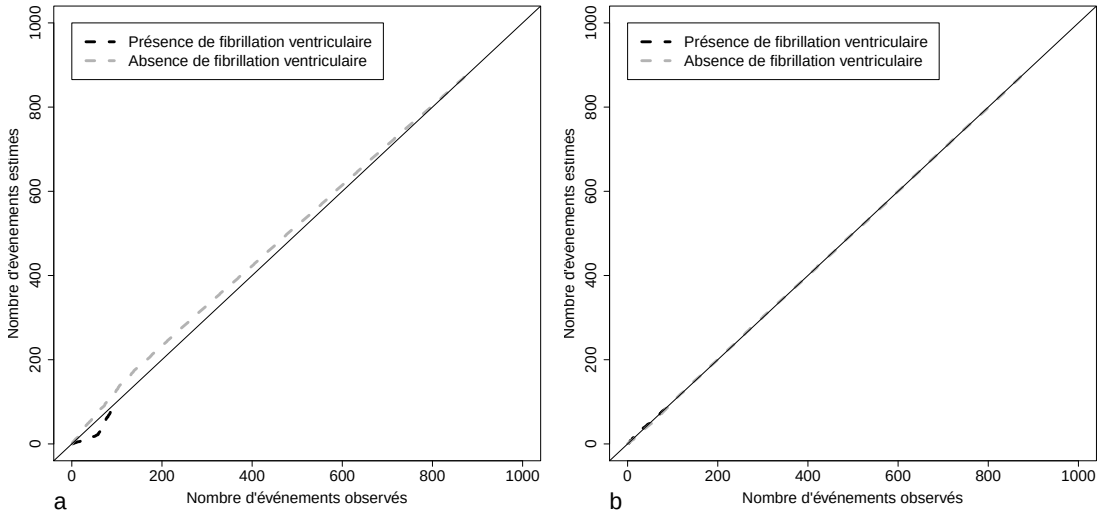


FIGURE 2.6. – Graphiques d'Arjas obtenus avec un modèle de Cox ajusté avec (a) la fibrillation ventriculaire sans effet et (b) la fibrillation ventriculaire avec un effet dépendant du temps

dant du temps a été ajouté au modèle multivarié et le test de corrélation des résidus de Schoenfeld devint statistiquement non-significatif pour toutes les covariables et globalement (tableau 2.2). Le modèle à risques instantanés multiplicatifs retenu fut donc le suivant :
$\lambda(t) = \lambda_0(t)e^{(\beta_{age}exp(age/100)\beta_{sexe}sexe+(\beta_{ic}+\beta_{ict}\times t)ic+(\beta_{dia}+\beta_{diat}t)dia+(\beta_{fv}+\beta_{fvt}t+\beta_{fvt2}(t-0{,}15)I(t>0{,}15))fv)}$.

Covariables	degré de significativité	
	modèle sans effet dépendant du temps pour le diabète	modèle final
exp(age/100)	0,130	0,099
sexe	0,622	0,695
ic	0,374	0,402
ict	0,218	0,241
dia	0,016	0,186
diat		0,230
fv	0,538	0,544
fvt	0,483	0,482
fvt2	0,479	0,478
global	0,160	0,571

Tableau 2.2. – Significativité des tests de corrélation des modèles à risques instantanés multivariés entre les résidus de Schoenfeld et les rangs des temps d'événement

Avec ce modèle multivarié à risques instantanés multiplicatifs, on constate que l'ensemble des paramètres étaient significativement différents de 0 (tableau 2.3) indiquant un effet significatif de chaque covariable sur le risque instantané de décès de base. L'exponentielle de l'âge et le sexe avaient un effet constant au cours du temps à la différence des autres covariables dont l'effet évoluait au cours du temps. Le rapport de risques instantanés associé au sexe était de 1,20 [1,05 ; 1,38] pour les femmes. On note qu'il était protecteur en analyse univariée (0,80 [0,70 ; 0,92]). En ce qui concerne l'âge, comme l'effet de l'exponentielle de l'âge divisé par 100 n'était pas linéaire, l'interprétation n'était pas aisée sans avoir recours à une représentation graphique. Celle-ci montre que l'effet de l'augmentation d'une année d'âge n'est pas la même quand on passait de 25 à 26 ans (1,038 [1,033 ; 1,043]) ou de 85 à 86 ans (1,070 [1,061 ; 1,079]) : le rapport des risques instantanés pour l'augmentation d'une année d'âge était plus important pour les patients âgés par rapport aux patients jeunes.

Les rapports des risques instantanés concernant les trois autres covariables évoluant au cours du temps, une représentation graphique de ceux-ci au cours du temps permet une meilleure interprétation (figure 2.9). Le rapport des risques instantanés associé à l'insuffisance cardiaque décroissait linéairement au cours du temps de 2,573 [2,055 ; 3,222] au début du suivi à 1,220 [0,798 ; 1,865] à la fin où il devenait non-significatif. Le rapport des risques instantanés associé au diabète, quant à lui, croissait linéairement

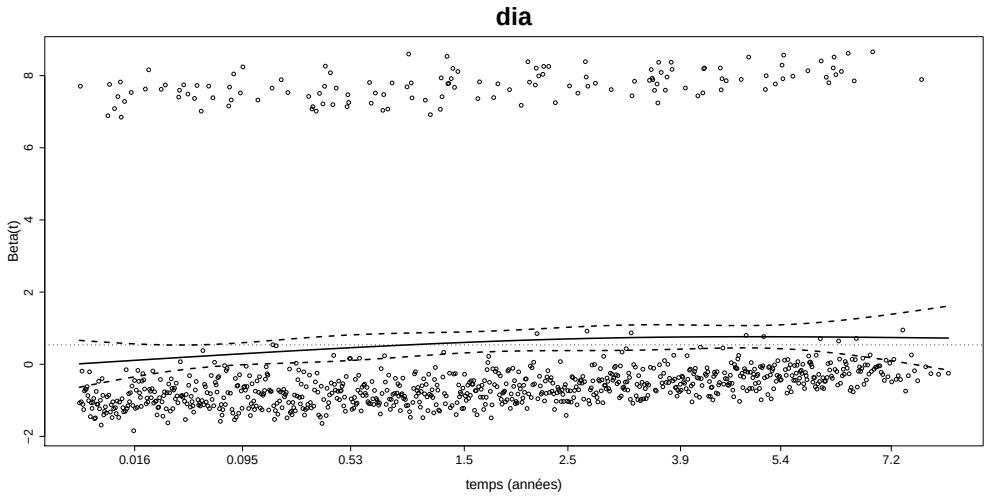


FIGURE 2.7. – Résidus de Schoenfeld en fonction du rang des événements pour le diabète ajusté avec un modèle à risques instantanés multiplicatifs multivarié. Une courbe de lissage en trait plein correspondant à l'estimation du paramètre β avec sa bande de confiance à 95 % en traits discontinus est ajoutée.

Covariable	$\hat{\beta}$	Rapport des risques instantanés ajustés $[IC_{95\%}]$	degré de significativité
exp(age/100)	2,87	17,56 [12,35 ; 24,97]	$< 2 \times 10^{-16}$
sexe	0,18	1,20 [1,05 ; 1,38]	0,00982
icf	0,95	2,57 [2,06 ; 3,22]	$<2 \times 10^{-16}$
ict	-0,08	0,92 [0,86 ; 0,98]	0,01024
dia	0,35	1,42 [1,09 ; 1,84]	0,00889
diat	0,09	1,10 [1,01 ; 1,19]	0,03137
fv	2,24	9,39 [6,32 ; 13,94]	$< 2 \times 10^{-16}$
fvt	-13,52	$1,35 \times 10^{-6}[1,03 \times 10^{-8}; 1,77 \times 10^{-4}]$	$5,61 \times 10^{-8}$
fvt2	13,49	$7,22 \times 10^{5}[4,93 \times 10^{3}; 1,06 \times 10^{8}]$	$1,15 \times 10^{-7}$

Tableau 2.3. – Paramètres $\hat{\beta}$ et rapports des risques instantanés ajustés avec l'intervalle de confiance à 95 % des neuf covariables prédictives analysées avec un modèle de Cox étendu multivarié.

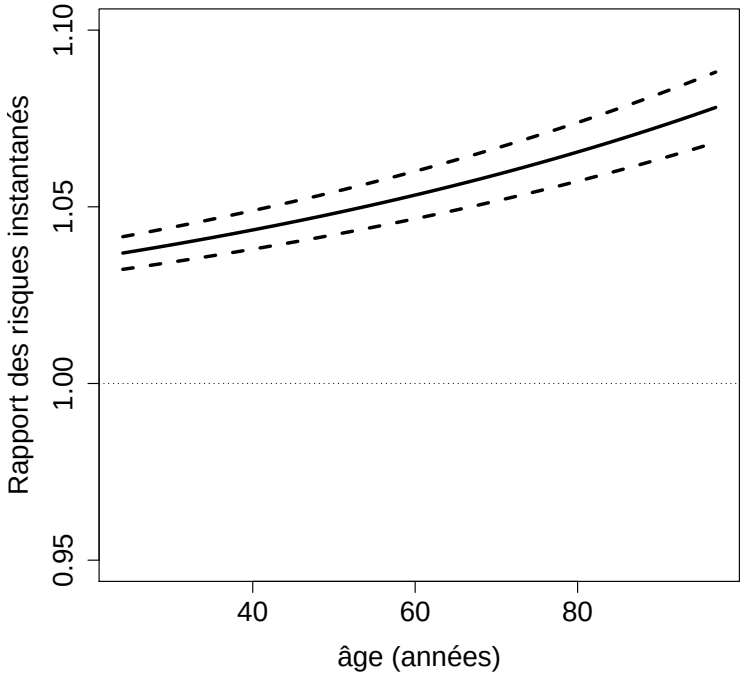


FIGURE 2.8. – Rapport des risques instantanés pour l'augmentation d'un an de l'âge en trait plein avec sa bande de confiance à 95 % en traits discontinus en fonction de l'âge.

au cours du temps de 1,415 [1,091 ; 1,835] au début du suivi à 3,241 [1,781 ; 5,897]. Enfin le rapport des risques instantanés associé à la fibrillation ventriculaire était très important au début (9,386 [6,318 ; 13,942]) avant de décroître très rapidement et de devenir non-significatif au bout de deux mois de suivi (1,235 [0,733 ; 2,082]).

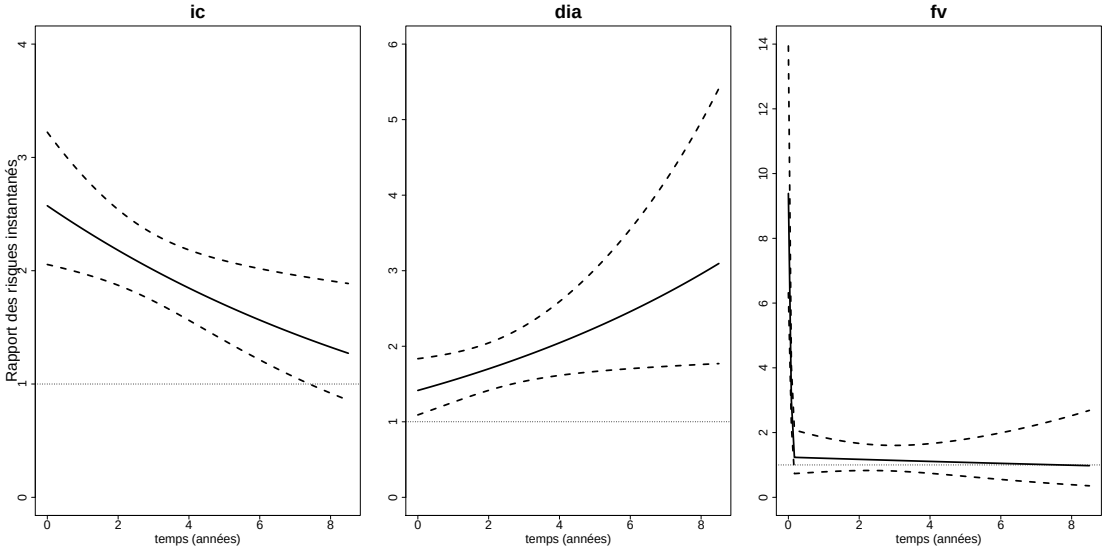


FIGURE 2.9. – Rapport des risques instantanés pour l'insuffisance cardiaque, le diabète et la fibrillation ventriculaire en fonction du temps.

2.4.2. Modélisation additive

2.4.2.1. Hypothèse de linéarité

Afin de définir la forme fonctionnelle de la covariable âge, il est nécessaire de la représenter graphiquement à l'aide de la courbe de lissage obtenue à partir de l'opposé du logarithme des pseudo-observations divisé par le temps en fonction de l'âge aux neuf déciles des temps d'événement. Ces courbes de lissage (figure 2.10) ne sont pas des droites, ce qui signifiait l'absence de linéarité mais sont presque parallèles ce qui signifiait une constance de l'effet au cours du temps. Les sept courbes représentant les déciles du troisième au neuvième sont très similaires. Elles sont en effet presque horizontales jusqu'à l'âge de 60 ans et croissent ensuite, ce qui indique que l'effet de l'âge était peu important jusqu'à 60 ans et qu'il augmentait pour les âges supérieurs. Aussi, cette forme correspond à peu près à une fonction exponentielle. L'augmentation est plus importante pour les deux premiers déciles mais les courbes ont également une forme exponentielle. L'âge a été divisé par 10 afin d'obtenir des valeurs de son exponentielle qui ne fussent pas trop grandes et un modèle d'Aalen fut ajusté avec l'exponentielle de l'âge divisé par 10.

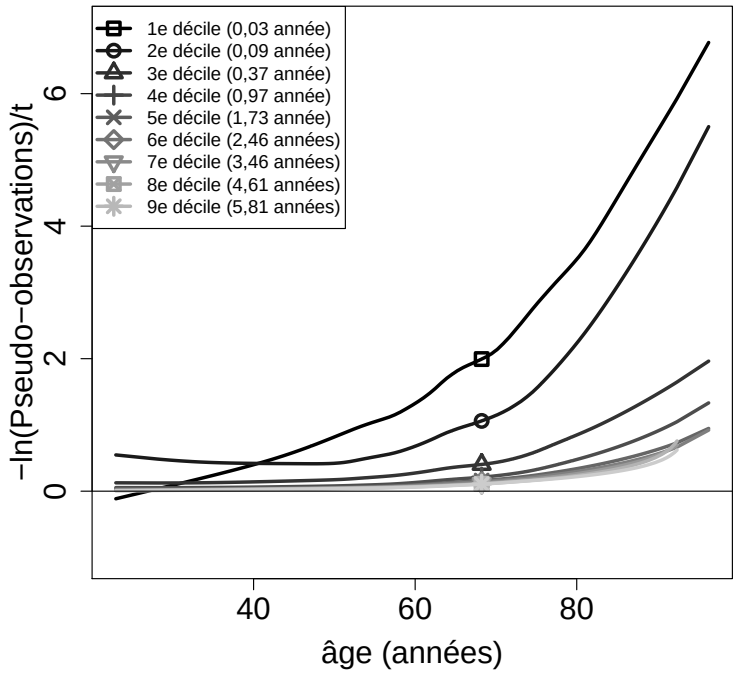


FIGURE 2.10. – Courbes de lissage obtenues à partir des pseudo-observations transformées en fonction de l'âge aux neuf déciles des temps d'événement (en années).

Pour s'assurer que la forme fonctionnelle retenue vérifie l'hypothèse de linéarité, les processus de résidus de martingale sont représentés graphiquement au cours du temps après avoir divisé l'échantillon en quatre strates définies par les quartiles de l'âge (figure 2.11).

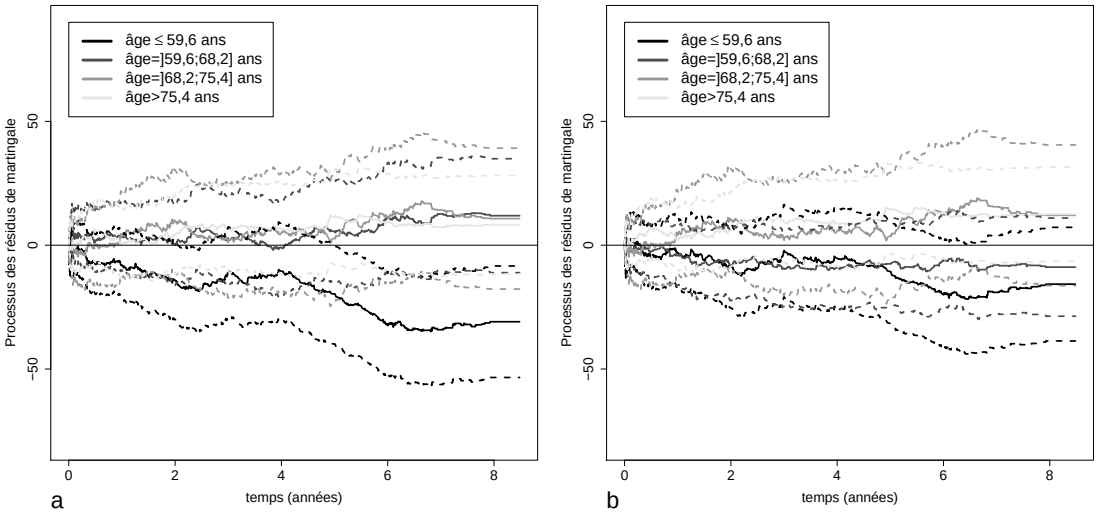


FIGURE 2.11. – Processus des résidus de martingale en trait plein avec leurs bandes de confiance en traits discontinus pour les quatre strates définies selon les quartiles de l'âge obtenus avec un modèle d'Aalen ajusté avec (a) l'exponentielle de l'âge divisé par 10 et (b) l'exponentielle de l'âge divisé par 10 avec un nœud à 70 ans.

Les courbes sont proches de l'axe des abscisses excepté celle concernant les patients les plus jeunes qui dévie significativement. Le modèle d'Aalen avec l'exponentielle de l'âge divisé par 10 surestimait la mortalité, les processus de résidus de martingale étant négatifs, notamment à la fin du suivi. Ceci fut confirmé par le résultat des tests du χ^2, le degré de significativité étant inférieur à 0,05 uniquement pour la strate des patients les plus jeunes. Il a donc été nécessaire de changer la modélisation de la forme fonctionnelle. Comme une autre fonction simple ne semblait pas correspondre à la forme fonctionnelle, l'ajout d'une covariable avec un nœud (exp(age-n)I(age>n)) permit de rendre plus flexible la fonction. Étant donné que l'augmentation commence à 70 ans, c'est ce seuil qui fut retenu comme nœud. Les courbes des processus de résidus de martingale sont alors proches de l'axe des abscisses et le résultat des tests du χ^2 montrait que ce modèle s'ajustait correctement aux données (tableau 2.4).

2.4.2.2. Hypothèse de la constance de l'effet

L'hypothèse de la constance des effets peut être vérifiée en représentant graphiquement les risques instantanés cumulés estimés avec le modèle d'Aalen et en ajoutant

Âge	degré de significativité	
	exp(age/10)	exp(age/10)+exp(age/10-7)I(age>70)
$\leq 59,6$	0,007	0,179
$(59,6-68,2]$	0,310	0,381
$(68,2-75,4]$	0,459	0,404
$>75,4$	0,416	0,197
global	0,056	0,221

Tableau 2.4. – Résultats des tests du χ^2 des processus de résidus de martingale pour la covariable âge ajustée avec un modèle à risques instantanés additifs univarié

ceux estimés avec le modèle de Lin afin de voir si ces derniers sont inclus dans les bandes de confiance des risques instantanés cumulés estimés avec le modèle d'Aalen. Dans ce cas on peut considérer que l'effet était constant au cours du temps. La représentation graphique de ces risques instantanés cumulés (figure 2.12) montre que l'on pouvait retenir l'hypothèse de la constance des effets pour l'exponentielle de l'âge divisé par 10, la covariable exp(age-70)I(age>70), le sexe et le diabète. En revanche, cette hypothèse n'était pas vérifiée pour l'insuffisance cardiaque et la fibrillation ventriculaire.

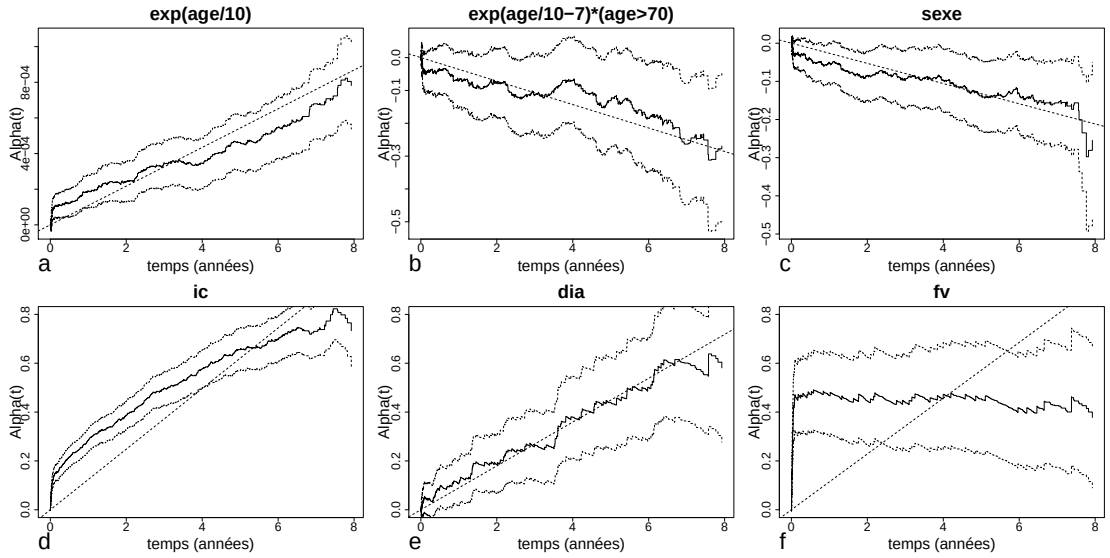


FIGURE 2.12. – Risques instantanés cumulés avec leurs bandes de confiance à 95 % en pointillés estimés avec un modèle d'Aalen pour les six covariables en analyse univariée. Les traits discontinus représentent les risques instantanés cumulés estimés avec le modèle de Lin en analyse multivariée.

On observe que les risques instantanés cumulés étaient négatifs pour le sexe et la covariable exp(age-70)I(age>70). Ceci se retrouve régulièrement et ne pose pas de problème tant que la somme des risques cumulés pour un individu reste supérieure à 0, sans quoi la survie serait alors supérieure à 1. Toutefois, si cette éventualité demeure possible théoriquement, on ne la retrouve que très rarement en pratique. On peut en conclure que le sexe féminin a un effet protecteur en analyse univariée.

2.4.2.3. Qualité de l'ajustement

Les graphiques d'Arjas obtenus (figure 2.13) après la division de l'âge en quatre strates définies sur ses quartiles montrent que la modélisation de l'âge de manière linéaire ne s'ajustait pas correctement pour chacune des strates. La modélisation de l'âge par son exponentielle après divison par 10 (figure 2.13) présentait un meilleur ajustement, toutefois, pour les patients les plus jeunes le nombre d'événements estimés par le modèle était supérieur au nombre d'événements observés, ce qui confirmait le résultat obtenu avec les processus de résidus de martingale. Les courbes obtenues avec le modèle avec l'exponentielle de l'âge divisé par 10 et la covariable exp(age-70)I(age>70) sont elles très proches de la diagonale confirmant un bon ajustement aux données (figure 2.14).

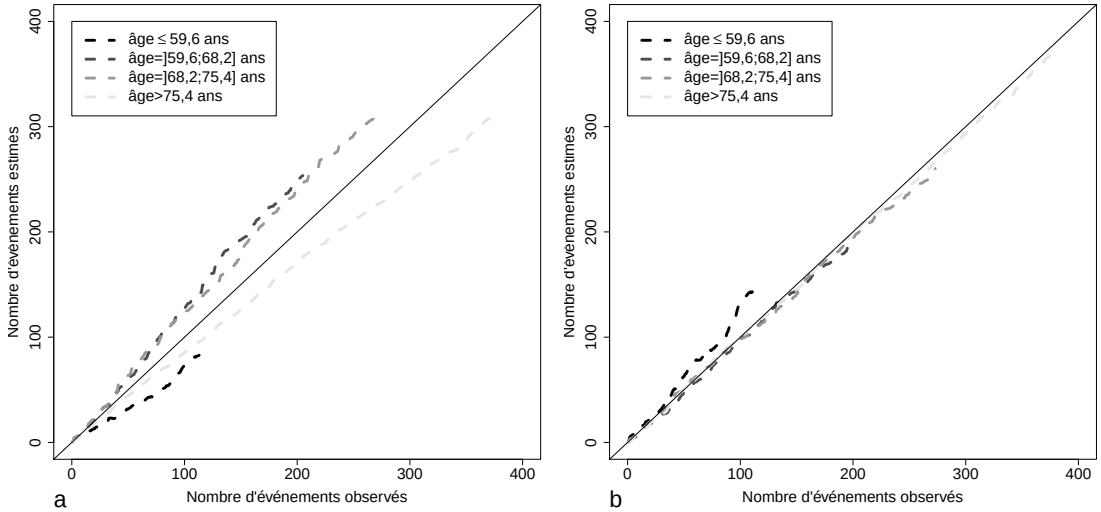


FIGURE 2.13. – Graphiques d'Arjas pour le modèle d'Aalen ajusté avec (a) l'âge sans transformation et (b) l'exponentielle de l'âge divisé par 10.

Ajustement du modèle à risques instantanés additifs multivarié

L'hypothèse de linéarité ayant été vérifiée pour chaque variable continue en analyse univariée, il restait encore à la vérifier en analyse multivariée en utilisant la même procédure. La représentation graphique des processus de résidus de martingale montre

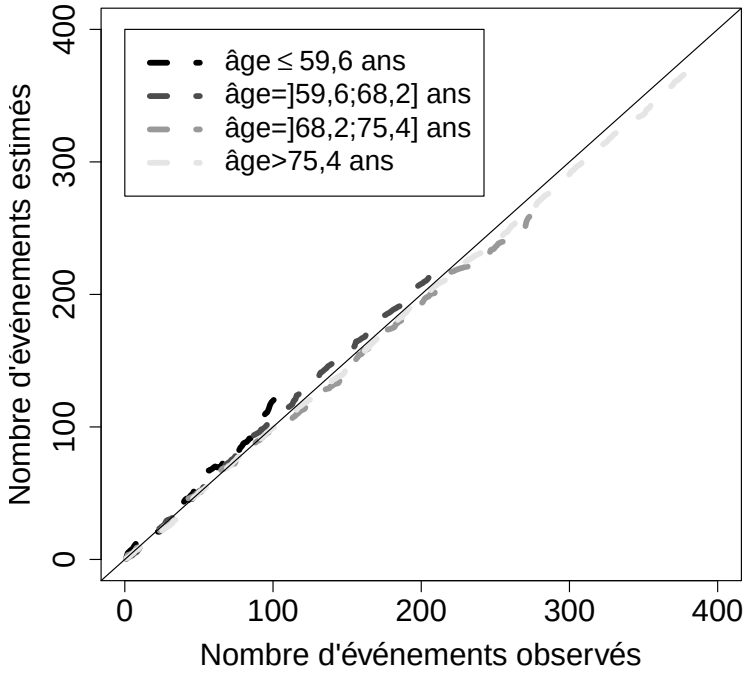


FIGURE 2.14. – Graphiques d'Arjas pour le modèle d'Aalen ajusté avec l'exponentielle de l'âge divisé par 10 et la covariable exp(age-70)I(age>70).

qu'ils ne s'éloignent pas significativement de l'axe des abscisses (figure 2.15) ce que confirmaient les tests du χ^2 de ces processus (tableau 2.5). Le modèle à risques intantanés additifs final retenu fut donc le suivant : $\lambda(t) = \alpha_0(t) + \alpha_{age1}(t)exp(age/10) + \alpha_{age2}(t)exp\big(age-70\big) \times I(age>70) + \alpha_{sexe} \times sexe + \alpha_{ic} \times ic + \alpha_{dia} \times dia + \alpha_{fv} \times fv$.

Âge	degré de significativité
$\leq 59,6$	0,542
$(59,6-68,2]$	0,481
$(68,2-75,4]$	0,757
$> 75,4$	0,316
global	0,589

Tableau 2.5. – Résultats des tests du χ^2 des processus de résidus de martingale pour la covariable âge ajustée avec un modèle à risques instantanés additifs avec les six covariables.

Les risques instantanés cumulés de ce modèle représentés sur la figure 2.16 étaient tous significativement différents de 0 (tableau 2.6) indiquant un effet significatif sur la mortalité.

Covariable	degré de significativité
ordonnée à l'origine	$1,20 \times 10^{-2}$
exp(age/10)	$2,85 \times 10^{-9}$
exp(age-70)I(age>70)	$3,60 \times 10^{-2}$
sexe	$8,58 \times 10^{-3}$
ic	$2,34 \times 10^{-23}$
dia	$2,31 \times 10^{-6}$
fv	$1,29 \times 10^{-7}$

Tableau 2.6. – Significativité des fonctions de régression du modèle d'Aalen des six covariables prédictives analysées avec un modèle d'Aalen étendu multivarié.

Les risques instantanés cumulés étaient positifs pour toutes les covariables binaires et leur pente était constante pour le sexe et le diabète mais plus importante avec cette dernière. Le risque de décès dû au diabète était donc supérieur à celui dû au sexe féminin. Par ailleurs, comme pour le modèle à risques instantanés multiplicatifs, l'effet protecteur du sexe féminin en analyse univariée devint un facteur de risque en analyse multivariée. Concernant l'insuffisance cardiaque, la pente diminue au cours du temps indiquant que son effet sur la mortalité diminuait également au cours du

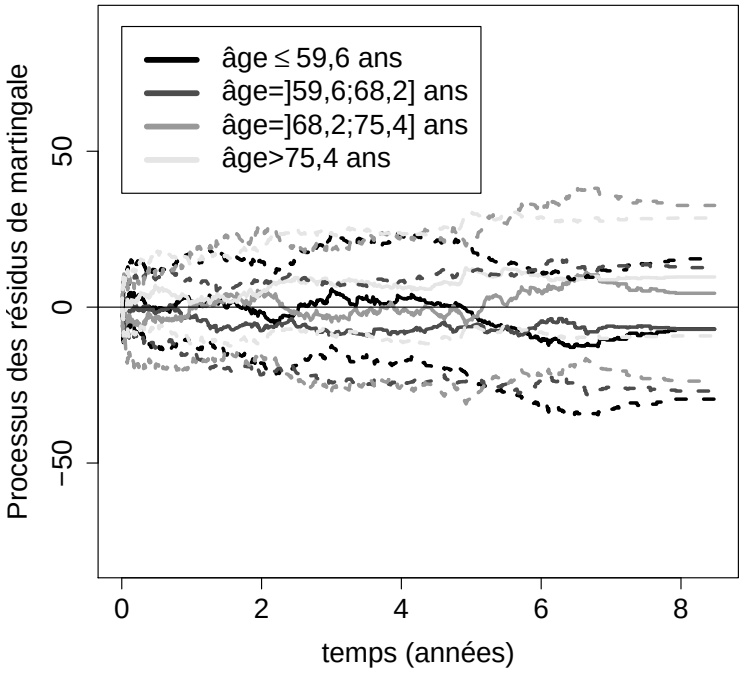


FIGURE 2.15. – Processus des résidus de martingale en trait plein avec leurs bandes de confiance en traits discontinus pour les quatre strates définies selon les quartiles de l'âge obtenus avec un modèle d'Aalen ajusté en analyse multivariée.

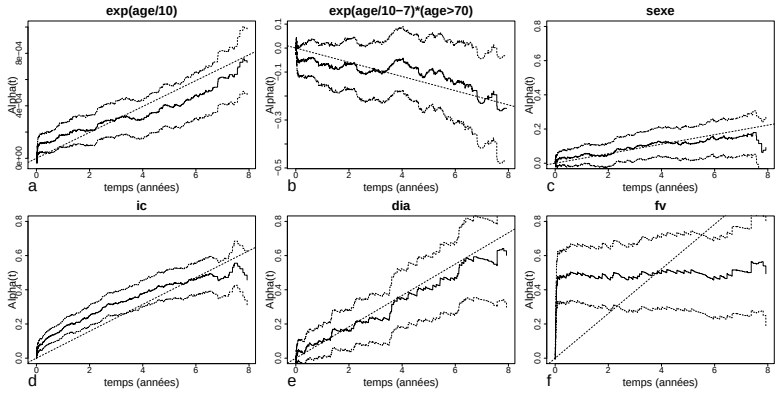


FIGURE 2.16. – Risques instantanés cumulés avec leurs bandes de confiance à 95 % en pointillés estimés avec un modèle d'Aalen pour les six covariables en analyse multivariée. Les traits discontinus représentent les risques instantanés cumulés estimés avec le modèle de Lin en analyse multivariée.

temps. Pour la fibrillation ventriculaire, la pente est très importante jusqu'à environ 0,1 an puis s'estompe totalement indiquant que le risque dû à cette covariable était très important au début du suivi et que, passé une quanrantaine de jours, le risque devenait nul. Finalement, pour l'âge, la pente est positive et constante pour l'exponentielle divisée par 10 et négative est constante pour la covariable exp(age-70)I(age>70). Par conséquent, on avait une augmentation exponentielle de l'effet de l'augmentation d'une année d'âge jusqu'à 70 ans puis une augmentation moins importante après 70 ans que ne l'aurait été une augmentation purement exponentielle (figure 2.17).

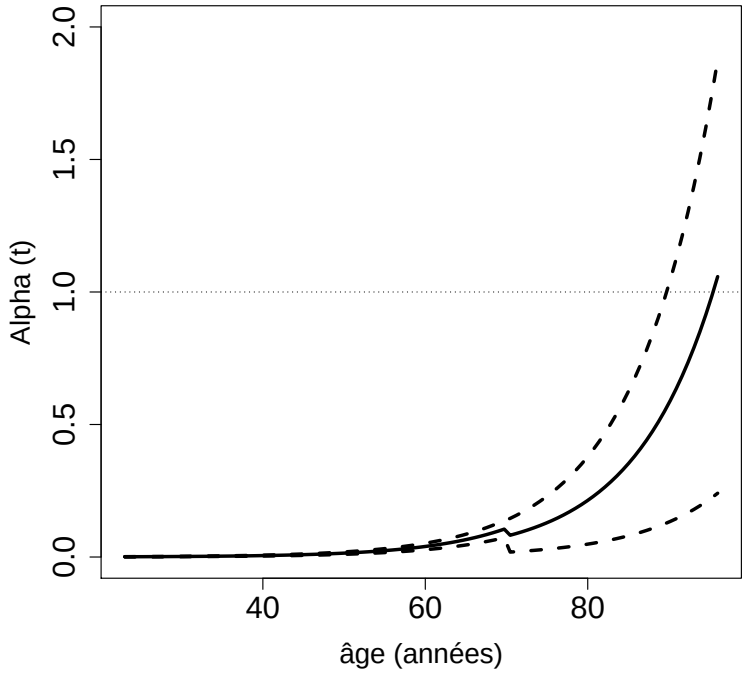


FIGURE 2.17. – Risques instantanés cumulés avec leur bande de confiance à 95 % en traits discontinus estimé avec le modèle multivarié de Lin pour l'augmentation d'une année d'âge.

2.5. Discussion

Ce travail propose une stratégie permettant de modéliser de manière optimale un modèle à risques instantanés multiplicatifs et additifs. Plusieurs points peuvent être développés. Tout d'abord, afin de modéliser correctement des données de survie, le plus important est de connaître le type d'effet des covariables sur le risque instantané afin de choisir le modèle à risques instantanés additifs ou multiplicatifs correspondant

à cet effet [12] et [2]. Il est toutefois difficile de connaître à l'avance ou théoriquement le type d'effet des covariables sur le risque instantané et il est donc souvent nécessaire de choisir un modèle en fonction d'autres caractéristiques, à savoir notamment la formulation des résultats, c'est-à-dire soit des rapports de risques instantanés interprétés comme des risques relatifs, soit des différences de risques instantantanés cumulés interprétées comme des risques instantanés cumulés dus à une covariable approximant des incidences quand ils sont petits. Les rapports de risques instantanés sont très souvent préférés car ils permettent de mesurer l'effet multiplicatif avec une seule valeur en cas de proportionnalité des risques et sont très répandus et utilisés dans les domaines médicaux et épidémiologiques. Les risques instantanés cumulés peuvent paraître plus complexes à interpréter notamment quand l'effet n'est pas constant ou qu'il est important mais ils représentent l'importance en valeur absolue du risque instantané dû à une covariable. Dans un contexte épidémiologique ou de prévention, ils apportent donc une information intéressante. Les modèles à risques instantanés additifs semblent meilleurs car ils permettent de représenter directement la variation de l'effet des covariables au cours du temps du fait de l'estimation non-paramétrique des fonctions de régression. En revanche, dans le cas des modèles à risques instantanés multiplicatifs, il est nécessaire de vérifier l'hypothèse de proportionnalité en utilisant les résidus de Schoenfeld. Quand cette hypothèse n'est pas vérifiée, un modèle de Cox étendu avec un effet dépendant du temps doit être réalisé. La stratégie proposée est intéressante car elle permet d'obtenir parallèlement un modèle à risques instantanés multiplicatifs et un modèle à risques instantanés additifs. On obtient alors des modèles dont les résultats apportent des informations différentes permettant une meilleure interprétation des données. Toutefois, elle ne permet pas directement de choisir pour chacune des covariables le type d'effet sur le risque instantané et de modéliser des modèles multiplicatifs et additifs comme le modèle de Cox-Aalen [43] et [14]. L'approche proposée au troisième chapitre pourrait éventuellement être utilisée pour développer une stratégie dans ce sens.

Deuxièmement, la stratégie proposée présente l'avantage d'être facilement réalisable d'autant qu'elle utilise des outils diagnostiques que l'on retrouve dans la plupart des logiciels de statistique. En effet, celle-ci utilise pour les modèles à risques instantanés multiplicatifs les résidus de Schoenfeld et de martingale qui sont bien connus mais aussi les pseudo-observations qui sont aisément calculables et les graphiques d'Arjas que l'on peut représenter facilement. D'autres approches ont été proposées. Dans celle de Sasieni et Winnett [42], reposant sur les résidus de la différence de martingale, une limite est la nécessité de définir précisément la forme fonctionnelle, même pour les grosses bases de données. Notre stratégie utilisant les pseudo-observations permet de représenter graphiquement non seulement la forme fonctionnelle des covariables continues mais aussi l'évolution au cours du temps de leurs effets. Ainsi, notre stratégie permet de retenir la meilleure forme fonctionnelle. Dans la stratégie proposée par Abrahamowicz et McKenzie [5], le modèle à risques instantanés multiplicatifs permet de modéliser directement des covariables continues ne respectant ni l'hypothèse de la proportionnalité des risques instantanés, ni celle de la log-linéarité. Comme ce modèle

est très flexible, il peut être sujet à un surajustement. Ce risque est limité avec notre approche car elle nécessite l'écriture explicite de la forme fonctionnelle dont l'équation peut parfois être délicate à trouver. Dans notre stratégie, les modèles à risques instantanés additifs sont modélisés en utilisant les pseudo-observations, les processus de résidus de martingale et les graphiques d'Arjas. D'autres outils ont été proposés pour ce type de modèles. Martinussen et Scheike [30] ont développé deux tests basés sur les processus gaussiens afin de tester la constance de l'effet. Contrairement à notre approche, dans laquelle la constance des effets est vérifiée graphiquement, celle de Martinussen et Scheike a trois limites principales. La première est la possibilité d'une discordance entre les résultats des deux tests, la seconde est le rejet de l'hypothèse nulle quand la taille de l'échantillon augmente même quand l'effet est faible et la troisième est l'absence de robustesse des tests implémentés dans leur progiciel `timereg`. Enfin, l'approche proposée par McKeague et Utikal [34] consiste à comparer un estimateur obtenu avec le modèle d'Aalen avec un estimateur non-paramétrique. La limite de cette approche est qu'elle nécessite des échantillons d'au moins 1000 sujets pour fonctionner correctement. De plus, les tests d'ajustement ne donnent pas la même information que les outils graphiques. Alors que les premiers mesurent quantitativement l'adéquation des modèles aux données, les seconds permettent de sélectionner les meilleures fonctions et formes fonctionnelles.

Troisièmement, la stratégie proposée donne des résultats qui sont proches de ceux obtenus par la stratégie employée par Martinussen et Scheike [32] pour analyser la même base de données. Dans notre stratégie, les résultats des modèles à risques instantanés multiplicatifs et additifs ne s'interpètent pas de la même façon. Si les effets des covariables sont significatifs avec les deux types de modèles, ils n'agissent pas de la même manière sur le risque instantané de base. Les covariables sexe, insuffisance cardiaque et fibrillation ventriculaire ont des effets qui sont respectivement constant, décroissant et important puis nul avec les deux types de modèle. Pour l'âge, les deux types de modèle montrent que l'effet est constant au cours du temps et qu'il augmente à partir de 70 ans. Cette augmentation est plus marquée quand elle est exprimée en termes de risques instantanés cumulés qu'en termes de rapport de risques instantanés. Le modèle à risques instantanés additifs souligne mieux l'importance de l'âge comme cause de la mortalité. En ce qui concerne le diabète, l'effet est constant avec le modèle à risques instantanés additifs mais croissant avec le modèle à risques instantanés multiplicatifs. Le modèle de Cox étendu montre ainsi que le risque relatif de décès croît au cours du temps pour les patients diabétiques, alors que le modèle à risques instantanés additifs indique que la mortalité due au diabète est constante au cours du temps. Cet écart s'explique par le fait que les patients diabétiques ont initialement un risque de mortalité inférieur par rapport aux patients, notamment ceux avec une fibrillation ventriculaire, ce qui entraîne un accroissement de la proportion de patients diabétiques au sein de l'échantillon au cours du temps. Par conséquent, le rapport de risques instantanés croît artificiellement alors que le risque de décès dû au diabète demeure constant. Tout ceci suggère que le modèle à risques instantanés additifs permet une meilleure interprétation des données et qu'il serait bon qu'il soit plus

souvent utilisé.

En conclusion, la stratégie proposée permet une analyse de la survie avec des modèles à risques instantanés multiplicatifs et additifs de manière optimale. Toutefois, elle ne permet pas le choix du meilleur modèle entre celui à risques instantanés multiplicatifs et celui à risques instantanés additifs. Cette question sera abordée dans le troisième chapitre.

Valorisation scientifique

Ce travail a été valorisé scientifiquement :

Publication scientifique :

François Lefebvre, Roch Giorgi (2021). « A strategy for optimal fitting of multiplicative and additive hazards regression models ». BMC medical research methodology, XX(X).(Soumis)

Communications orales :

dans des séminaires internes :

« Intérêt de l'étude des modèles additifs pour l'analyse de survie », séminaire interne du SESSTIM, Marseille, 09/06/2017

https://youtu.be/6akwl_bv1j0

Communications affichées :

dans des conférences internationales à comité scientifique de sélection :

« Exploring and understanding survival data using Cox proportional hazards and Aalen's additive model ». Lefebvre François, Giorgi Roch and the working survival group CENSUR. 37th annual conference of International Society for Clinical Biostatistics, Birmingham, Royaume-uni

Programme R :

disponible en annexe et à venir sur le site du SESSTIM

3. Comparaison des modèles à risques instantanés multiplicatifs et additifs

Sommaire

3.1 Pseudo-résidus . . . 57
3.1.1 Définition . . . 58
3.1.2 Somme des carrés des pseudo-résidus . . . 58
3.2 Étude de simulations . . . 59
3.2.1 Une variable . . . 59
3.2.2 Deux variables . . . 59
3.3 Analyses des simulations . . . 60
3.3.1 Une seule variable continue . . . 61
3.3.2 Une seule variable binaire . . . 63
3.3.3 Deux covariables . . . 65
3.4 Applications . . . 74
3.4.1 Cirrhose biliaire primitive . . . 74
3.4.2 Cancer du sein . . . 75
3.4.3 Infarctus du myocarde . . . 76
3.5 Discussion . . . 76

3.1. Pseudo-résidus

Afin de sélectionner le modèle s'ajustant le mieux aux données entre un modèle à risques instantanés multiplicatifs et un modèle à risques instantanés additifs, différentes approches ont été proposées. En effet, ces modèles n'étant pas emboîtés, il n'est pas possible de les comparer simplement à l'aide d'un test. De plus, des critères tels que l'AIC (critère d'information d'Akaike) ou le BIC (critère d'information bayésien) sont difficilement utilisables, car si la vraisemblance totale du modèle de Cox est estimable tout comme celle du modèle d'Aalen [28], il n'est pas possible de définir simplement le nombre de paramètres du modèle d'Aalen, les estimations étant des fonctions de régression. De plus, dans le cadre de la survie, certains sujets étant censurés, il n'existe pas d'équivalent aux résidus du modèle linéaire. Il existe toutefois des résidus dans le cadre de l'analyse de survie comme les résidus de martingale

et de Schoenfeld. Les résidus de martingale sont un peu similaires aux résidus du modèle linéaire, étant donné qu'ils représentent une différence entre une donnée observée, le processus de comptage et une donnée estimée, le compensateur. Comme le compensateur est une fonction croissante au cours du temps définie sur $[0;+\infty[$, les résidus de martingale diminuent au cours du temps. Tout critère basé sur ces résidus a pour conséquence une augmentation artificielle du poids des sujets au cours du temps. Aussi, une solution basée sur un autre type de résidus, les pseudo-résidus, a été proposée et étudiée dans ce chapitre.

3.1.1. Définition

Pohar-Perme et Andersen ont proposé un nouveau type de résidus, nommés pseudo-résidus, basés sur les pseudo-observations [39] afin de vérifier l'ajustement des modèles de Cox et de Lin aux données. Ces pseudo-résidus sont définis comme la différence entre une pseudo-observation $\hat{S}_i(t)$ et une estimation de la survie obtenue par un modèle $\hat{S}(t|x_i)$ et s'écrit :

$$\hat{\epsilon}_i(t) = \hat{S}_i(t) - \hat{S}(t|x_i) = n\hat{S}(t) - (n-1)\hat{S}^{-i}(t) - \hat{S}(t_s|x_i). \tag{3.1}$$

3.1.2. Somme des carrés des pseudo-résidus

Ces pseudos-résidus sont analogues aux résidus des modèles de régression linéaire car ils représentent une différence entre une pseudo-observation de la survie et une estimation de la survie obtenue à partir d'un modèle. Dans le cadre des modèles linéaires, les fonctions de régression sont estimées par la méthode des moindres carrés ordinaires, c'est-à-dire par la minimisation de la somme des carrés des résidus qui représente la somme des carrés des distances entre une valeur observée et une valeur estimée pour chaque sujet. Dans le cadre de la survie, il est donc possible de remplacer la valeur observée par la pseudo-observation et la valeur estimée par celle estimée par le modèle à risques instantanés additifs, multiplicatifs ou simultanément multiplicatifs et additifs. On obtient ainsi différents types de pseudo-résidus selon le modèle utilisé : $\hat{\epsilon}_i(t)_{mult} = \hat{S}_i(t) - \hat{S}_{mult}(t|x_i)$ avec un modèle à risques instantanés multiplicatifs, $\hat{\epsilon}_i(t)_{add} = \hat{S}_i(t) - \hat{S}_{add}(t|x_i)$ avec un modèle à risques instantanés additifs et $\hat{\epsilon}_i(t)_{m-a} = \hat{S}_i(t) - \hat{S}_{m-a}(t|x_i)$ avec un modèle à risques instantanés simultanément multiplicatifs et additifs. Afin de sélectionner le modèle qui s'ajuste le mieux aux données, s'inspirant de la statistique du PRESS (somme des carrés des erreurs résiduelles prédites), nous avons proposé d'utiliser la somme des carrés des pseudo-résidus $\sum_{i=1}^{n} \hat{\epsilon}_i^2(t)$ (SCPR) comme méthode de sélection entre un modèle à risques instantanés multiplicatifs, additifs et simultanément multiplicatifs et additifs en retenant celui qui a la plus petite somme. Il faut noter que puisque cette somme est une somme de carrés des distances pour chaque sujet, celle-ci croît avec la taille de l'échantillon. Comme pour la vraisemblance, sa valeur n'a donc pas de signification intrinsèque et ne peut être interpétée qu'en fonction de la valeur d'une autre estimation basée sur les mêmes

données.

3.2. Étude de simulations

Afin d'évaluer les performances du critère que nous proposons, la SCPR la plus petite, pour sélectionner le modèle à risques instantanés additifs, multiplicatifs ou simultanément multiplicatifs et additifs, une étude de simulations a été réalisée.

3.2.1. Une variable

Dans un premier temps, des données de survie ont été générées avec une seule variable, selon deux scénarios, à savoir le scénario A avec un effet multiplicatif de la variable sur le risque instantané de base et le scénario B avec un effet additif de la variable sur le risque instantané de base. Celui-ci a été généré avec une loi de Weibull exponentialisée [36] qui permet une grande souplesse dans la génération des risques instantanés. Les modèles de génération étaient :

1. avec une covariable dont l'effet est constant : $\lambda_{A1}(t|z) = \lambda_0(t)e^{\beta_z z}$ et $\lambda_{B1}(t|z) = \alpha_0(t) + \alpha_z z$
2. avec une covariable dont l'effet dépend du temps de manière linéaire : $\lambda_{A2}(t|z) = \lambda_0(t)e^{(\beta_z+\beta_{tz}t)z}$ et $\lambda_{B2}(t|z) = \alpha_0(t) + (\alpha_z + \alpha_{tz}t)\,z$

Les modèles de génération avec un effet constant correspondent respectivement à un modèle de Cox et à un modèle de Lin, et ceux avec un effet dépendant du temps à un modèle de Cox avec un effet dépendant du temps et à un modèle d'Aalen. Les valeurs des paramètres ont été choisies afin d'obtenir trois tailles d'effet (faible, moyen et fort) et de telle sorte que l'effet de la variable obtenu avec une génération multiplicative soit similaire à celui de variables obtenu avec une génération additive, c'est-à-dire que la survie obtenue avec un effet multiplicatif soit similaire à celle obtenue avec un effet additif pour une valeur de la variable donnée. Pour chaque scénario, la variable était premièrement continue et suivait une distribution normale de moyenne 60 et d'écart-type 15, puis secondairement binaire avec la moitié des sujets par modalité. Une censure était générée à l'aide d'une loi uniforme permettant dans un premier temps d'obtenir un poucentage de 10 % de sujets censurés puis dans un deuxième temps 30 % sur quatre scénarios. Une censure administrative était appliquée à 10 ans de suivi. Pour chaque scénario, 1 000 simulations étaient réalisées avec trois tailles d'échantillon différentes, à savoir 500, 1 000 et 2 000.

3.2.2. Deux variables

Dans un second temps, des données de survie ont été générées avec deux covariables, une qualitative binaire z_1 et une quantitative z_2 selon quatre scénarios :

1. A3 : les deux covariables avaient simultanément un effet multiplicatif sur le risque instantané de base : $\lambda_{M3}(t|z_1, z_2) = \lambda_0(t)e^{(\beta_{z_1} z_1 + \beta_{z_2} z_2)}$

2. B3 : les deux covariables avaient simultanément un effet additif sur le risque instantané de base : $\lambda_{A3}(t|z_1,z_2) = \alpha_0(t) + \alpha_{z_1} z_1 + \alpha_{z_2} z_2$
3. C1 : la variable qualitative avait un effet additif et la variable quantitative un effet multiplicatif sur le risque instantané de base : $\lambda_{C1}(t|z_1,z_2) = \left(\lambda_0(t) + \alpha_{z_1} z_1\right) e^{(\beta_{z_2} z_2)}$
4. C2 : la variable qualitative avait un effet multiplicatif et la variable quantitative un effet additif sur le risque instantané de base : $\lambda_{C2}(t|z_1,z_2) = \left(\lambda_0(t) + \alpha_{z_2} z_2\right) e^{(\beta_{z_1} z_1)}$

Les modèles utilisés pour analyser les données simulées avec une seule variable étaient :

1. le modèle de Cox (correspondant au scénario de génération A1)
2. le modèle de Cox avec un effet dépendant du temps de manière linéaire (Coxdt) (correspondant au scénario de génération A2)
3. le modèle de Lin (correspondant au scénario de génération B1)
4. le modèle d'Aalen (correspondant au scénario de génération B2)

Les modèles utilisés pour analyser les données simulées avec deux variables étaient les mêmes que ceux utilisés pour analyser une seule variable avec en plus deux modèles de Cox-Aalen, c'est-à-dire :

1. le modèle de Cox (correspondant au scénario de génération A3)
2. le modèle de Cox avec un effet dépendant du temps de manière linéaire
3. le modèle de Lin (correspondant au scénario de génération B3)
4. le modèle d'Aalen
5. le modèle de Cox-Aalen avec un effet qualitatif additif et un effet quantitatif multiplicatif sur le risque instantané de base (modèle CA1 correspondant au scénario de génération C1)
6. le modèle de Cox-Aalen avec un effet qualitatif multiplicatif et un effet quantitatif additif sur le risque instantané de base (modèle CA2 correspondant au scénario de génération C2)

3.3. Analyses des simulations

Pour chaque figure, la différence des sommes des carrés des pseudo-résidus entre deux modèles m_1 et m_2 a été calculée selon la formule $(d_{(1-2)} = \sum_(i = 1)^n \hat{\epsilon}_i(t)_{m_1} - \sum_(i = 1)^n \hat{\epsilon}_i(t)_{m_2}$ et est représentée en bleu quand la génération est multiplicative, en rose quand elle est additive, en jaune quand elle est réalisée selon le modèle de Cox-Aalen de type 1, c'est-à-dire avec un effet additif de la covariable binaire et multiplicatif de la covariable continue et en rouge quand elle est réalisée selon le modèle de Cox-Aalen de type 2, c'est-à-dire avec un effet additif de la covariable continue et multiplicatif de la covariable binaire.

3.3.1. Une seule variable continue

La figure 3.1 montre les distributions des différences deux à deux des SCPR calculées avec les quatre modèles à risques instantanés multiplicatifs ou additifs (modèle de Cox, modèle de Cox avec un effet dépendant du temps linéaire, modèle de Lin et modèle d'Aalen) quand les données étaient générées avec une covariable continue avec un effet constant et moyen et un échantillon de 1 000 sujets, correspondant aux scénarios A1 et B1. Les différences des SCPR entre un modèle additif et multiplicatif étaient négatives dans environ 70 % des cas quand la génération des données était additive et positives dans environ 70 % des cas quand la génération des données était multiplicative. Par conséquent le modèle retenu était du même type (additif ou multiplicatif) que le modèle de génération dans environ 70 % des cas. Les différences des SCPR entre deux modèles additifs et deux modèles multiplicatifs étaient beaucoup plus petites que celles observées entre un modèle additif et un modèle multiplicatif quelle que soit la génération des données indiquant qu'il est difficile avec cet outil de différencier deux modèles additifs ou deux modèles multiplicatifs. Toutefois, elles étaient la plupart du temps négatives et donc en faveur du modèle le plus souple (Coxdt par rapport à Cox et Aalen par rapport à Lin).

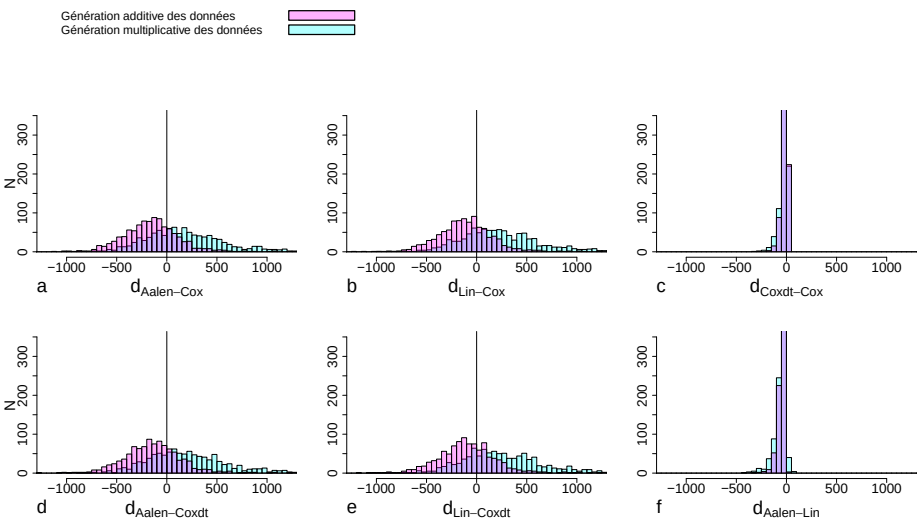


FIGURE 3.1. – Distributions des différences des SCPR estimées obtenues avec les quatre modèles à risques instantanés multiplicatifs ou additifs comparés deux à deux. Situation avec une variable générée continue constante et d'effet moyen et un échantillon de 1 000 sujets.

Quand la taille de l'échantillon augmentait (de 500 à 2 000 sujets) ou quand l'importance de l'effet augmentait, le pourcentage de fois où le modèle additif était retenu

quand le modèle de génération était additif augmentait jusqu'à 80 % et le pourcentage de fois où le modèle multiplicatif était retenu quand le modèle de génération était multiplicatif augmentait jusqu'à 90 % (tableaux 3.1 et 3.2).

Taille de l'échantillon	Taille de l'effet											
	Faible				Moyen				Fort			
Scénario	A1	B1	A2	B2	A1	B1	A2	B2	A1	B1	A2	B2
N=500	44,7	24,9	47,0	26,4	57,6	28,5	63,4	25,3	70,3	27,7	79,4	21,4
N=1000	53,2	29,1	54,6	24,6	66,6	25,6	73,7	21,6	77,4	24,6	90,1	20,9
N=2000	60,8	33,5	65,9	19,7	72,9	22,1	83,3	15,0	89,4	20,5	96,5	15,4

Tableau 3.1. – Pourcentage des cas pour lesquels le modèle de Cox est retenu par rapport au modèle d'Aalen en fonction du scénario de génération des données (A ou B ; 1 : quand la variable a un effet constant et 2 : quand la variable a un effet dépendant du temps linéaire), la taille de l'échantillon (500, 1 000, ou 2 000), et la taille de l'effet (faible, moyen, ou fort). Situations avec une seule variable continue générée

En revanche, la proportion de fois où l'on retenait le modèle de Coxdt par rapport au modèle de Cox (environ 80 %) ou le modèle d'Aalen par rapport au modèle de Lin (environ 90 %) ne variait pas quel que soit le modèle de génération, la taille de l'échantillon ou la taille de l'effet.

Taille de l'échantillon	Taille de l'effet											
	Faible				Moyen				Fort			
Scénario	A1	B1	A2	B2	A1	B1	A2	B2	A1	B1	A2	B2
N=500	57,5	42,5	58,6	46,7	67,0	40,5	69,9	41,4	74,0	35,8	82,5	42,0
N=1000	61,0	42,0	72,1	40,3	71,3	30,6	77,1	35,7	80,0	30,6	91,3	40,9
N=2000	65,7	40,6	80,6	40,5	76,0	25,4	85,0	27,9	91,0	24,2	96,7	35,8

Tableau 3.2. – Pourcentage des cas pour lesquels le modèle de Cox est retenu par rapport au modèle de Lin en fonction du scénario de génération des données (A ou B ; 1 : quand la variable a un effet constant et 2 : quand la variable a un effet dépendant du temps linéaire), la taille de l'échantillon (500, 1 000, ou 2 000), et la taille de l'effet (faible, moyen, ou fort). Situations avec une seule variable continue générée

L'ajout d'un seuil n'a pas permis de mieux discriminer les modèles à risques instantanés multiplicatifs et additifs quel que soit le modèle de génération (figure .5). Les

résultats obtenus avec une covariable continue ayant un effet dépendant du temps (scénarios A2 et B2), tableaux(3.1 et 3.2) étaient similaires à ceux obtenus avec une covariable dont l'effet était constant à l'exception de la comparaison entre le modèle d'Aalen et celui de Lin. En effet, avec un modèle de génération multiplicatif, quand la taille de l'effet et de l'échantillon augmentait, le modèle de Lin était plus souvent retenu que celui d'Aalen (jusqu'à 75 %). L'augmentation de la proportion de censures à 30 % ne modifiait pas fondamentalement la distribution des SCPR (figure .2).

3.3.2. Une seule variable binaire

En présence d'une seule variable binaire ayant un effet constant, les différences de SCPR entre un modèle d'Aalen et l'un des trois autres étaient pratiquement toujours négatives (figure 3.2). Par conséquent, le modèle d'Aalen était presque toujours retenu quel que soit le modèle de génération des données. Les différences des SCPR estimées par un modèle de Coxdt et par un modèle de Cox ou un modèle de Lin étaient souvent négatives ce qui entraînait la sélection du modèle de Coxdt dans plus de 80 % des cas quel que soit le modèle de génération. Les différences des SCPR estimées par un modèle de Lin et par un modèle de Cox étaient négatives dans 69,5 % des cas quand le modèle de génération était additif et dans 5,5 % des cas quand il était multiplicatif. Le modèle de Lin était donc retenu dans 69,5 % des cas et le modèle de Cox dans 94,5 % des cas, respectivement.

Quand la taille de l'échantillon augmentait (de 500 à 2 000 sujets) ou quand l'importance de l'effet augmentait, le pourcentage de fois où le modèle d'Aalen était retenu par rapport au modèle de Cox ne variait que très faiblement et restait très proche de 100 % (tableau 3.3). Quand le modèle de génération était multiplicatif, la proportion des cas pour lesquels le modèle de Cox était retenu par rapport au modèle de Lin augmentait avec la taille de l'effet ou de l'échantillon (tableau 3.4). Quand le modèle de génération était additif, seule l'augmentation de la taille de l'échantillon augmentait la proportion de cas pour lesquels le modèle de Lin était retenu par rapport au modèle de Cox. Afin d'augmenter cette proportion quand la taille de l'effet augmentait, il était nécessaire d'ajouter un seuil. L'augmentation de la valeur du seuil à 100 augmentait la proportion de cas pour lesquels le modèle approprié était retenu sauf pour la comparaison des modèles d'Aalen et de Coxdt. Dans ce cas, le changement de seuil n'améliorait pas la sélection (figure .4). Les résultats étaient similaires quand l'effet était linéairement dépendant du temps (tableaux 3.3 et 3.4). Étonnament, la proportion de cas pour lesquels le modèle de Lin était retenu par rapport au modèle de Cox décroissait quand la génération était additive. Le modèle de Lin s'ajustait mieux que le modèle de Cox avec des données générées selon un modèle multiplicatif par rapport à des données générées selon un modèle additif quand l'effet était linéairement dépendant du temps. L'augmentation de la proportion de censures à 30 % ne modifiait pas fondamentalement la distribution des SCPR (figure .3).

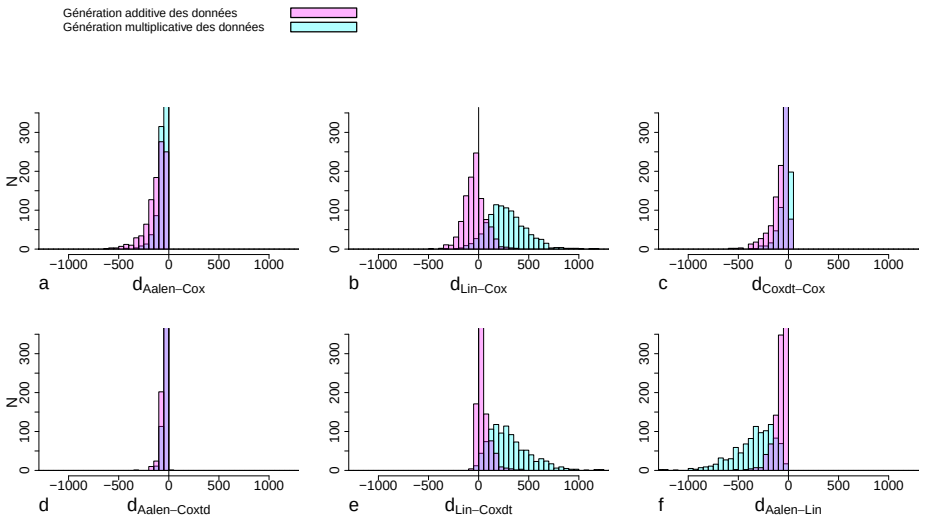


FIGURE 3.2. – Distributions des différences des SCPR estimées obtenues avec les quatre modèles à risques instantanés multiplicatifs ou additifs comparés deux à deux. Situation avec une variable générée binaire constante et d'effet moyen et un échantillon de 1 000 sujets.

Taille de l'échantillon	Taille de l'effet											
	Faible				Moyen				Fort			
Scénario	A1	B1	A2	B2	A1	B1	A2	B2	A1	B1	A2	B2
N=500	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,6	0,1	0,0	0,0
N=1000	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,5	2,7	0,2	0,2
N=2000	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	10,5	4,7	1,4	2,4

Tableau 3.3. – Pourcentage de cas pour lesquels le modèle de Cox est retenu par rapport au modèle d'Aalen en fonction du scénario de génération des données (A ou B ; 1 : quand la variable a un effet constant et 2 : quand la variable a un effet dépendant du temps linéaire), la taille de l'échantillon (500, 1 000, ou 2 000), et la taille de l'effet (faible, moyen, ou fort). Situations avec une seule variable binaire générée.

Taille de l'échantillon	Taille de l'effet											
	Faible				Moyen				Fort			
Scénario	A1	B1	A2	B2	A1	B1	A2	B2	A1	B1	A2	B2
N=500	84,0	22,8	63,0	76,7	86,9	42,0	69,1	85,4	96,4	61,1	83,0	89,7
N=1000	93,0	13,1	58,8	83,5	94,5	30,5	68,6	92,2	99,1	49,3	87,8	92,2
N=2000	97,9	5,5	54,6	90,3	98,3	21,2	66,5	96,7	100,0	40,0	94,1	98,1

Tableau 3.4. – Pourcentage de cas pour lesquels le modèle de Cox est retenu par rapport au modèle de Lin en fonction du scénario de génération des données (A ou B ; 1 : quand la variable a un effet constant et 2 : quand la variable a un effet dépendant du temps linéaire), la taille de l'échantillon (500, 1 000, ou 2 000), et la taille de l'effet (faible, moyen, ou fort). Situations avec une seule variable binaire générée.

3.3.3. Deux covariables

La figure 3.3 montre les distributions des différences de SCPR entre les quatre modèles à risques instantanés additifs ou multiplicatifs comparés deux à deux avec deux variables générées constantes et moyennes, l'une étant binaire et l'autre continue. Quand le modèle de génération était multiplicatif, les différences de SCPR entre les modèles à risques instantanés additifs et multiplicatifs étaient positives tout le temps et ceci jusqu'à un seuil de 2 300 représenté par le segment vertical rouge. Quand ce seuil dépassait 2 300, la proportion des cas pour lesquels le modèle retenu était additif croissait doucement. Quand le modèle de génération était additif, les différences des SCPR entre les modèles à risques instantanés additifs et multiplicatifs étaient positives dans 73 % des cas. Le modèle additif était donc retenu dans 27 % des cas. L'ajout d'un seuil permettait d'augmenter cette proportion qui atteignait 90 % pour une valeur du seuil de 2 300. La figure 3.4 montre que les résultats étaient très dépendants de la valeur du seuil. Une valeur du seuil de 4 000 semblait être celle qui permettait de retenir le plus souvent le modèle correspondant à la génération des données, c'est-à-dire un modèle additif quand la génération était additive et multiplicatif quand la génération était multiplicative.

Les distributions des différences de SCPR estimées par le modèle d'Aalen et les modèles de Cox-Aalen de type 1 et 2 étaient semblables à celle obtenue avec les modèles d'Aalen et de Cox, que la génération soit additive ou multiplicative (figure 3.5). Les modèles de Cox-Aalen de type 1 et 2 se comportaient donc comme des modèles de Cox en termes de pseudo-résidus. Les différences de SCPR entre les modèles de Cox-Aalen de type 1 ou 2 et le modèle de Cox étaient petites, notamment entre le modèle de Cox-Aalen de type 1 et le modèle de Cox. Ils s'ajustaient de manière équivalente aux données. Les différences de SCPR entre le modèle de Cox-Aalen de type 1 et celui

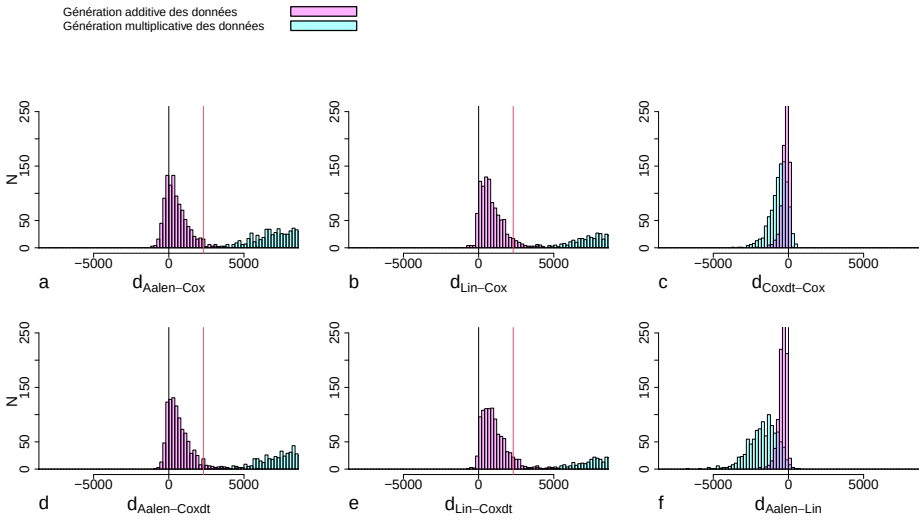


FIGURE 3.3. – Distributions des différences des SCPR estimées obtenues avec les quatre modèles à risques instantanés multiplicatifs ou additifs comparés deux à deux. Situation avec deux covariables générées constantes et moyennes, une binaire et une continue, et un échantillon de 2 000 sujets.

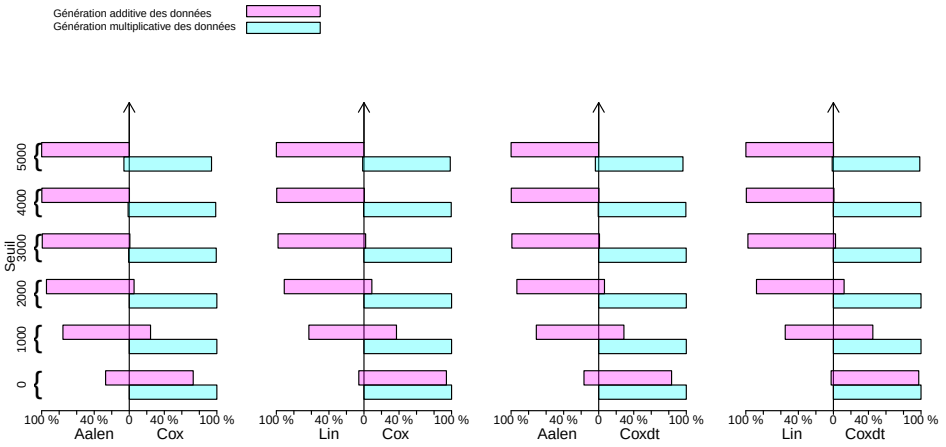


FIGURE 3.4. – Pourcentage de cas pour lesquels le meilleur modèle est retenu en fonction de la valeur du seuil et du modèle de génération des données entre les modèles à risques instantanés additifs (Aalen, Lin) et multiplicatifs (Cox, Coxdt). Situation avec deux covariables générées constantes et moyennes, une binaire et une continue, et un échantillon de 2 000 sujets.

de type 2 étaient également faibles, notamment quand la génération était additive.

Les distributions des différences de SCPR estimées par les modèles de Lin ou de Coxdt et les modèles de Cox-Aalen de type 1 et 2 étaient semblables à celles obtenues avec les modèles de Lin ou de Coxdt et le modèle de Cox, que la génération soit additive ou multiplicative (figure 3.6). De la même façon qu'avec le modèle d'Aalen, les modèles de Cox-Aalen se comportaient comme des modèles de Cox en termes de pseudo-résidus.

La figure 3.7 montre les distributions des différences de SCPR entre les quatre modèles à risques instantanés additifs ou multiplicatifs comparés deux à deux avec deux covariables générées, l'une binaire et l'autre continue, constantes et moyennes, l'une ayant un effet additif, l'autre un effet multiplicatif.

La comparaison entre les modèles à risques instantanés additifs et multiplicatifs montrait que les différences de SCPR étaient constamment positives, entraînant la sélection du modèle à risques instantanés multiplicatifs dans la totalité des cas. L'ajout d'un seuil ayant pour valeur 7 000 (représenté par un segment vertical rouge) permettait de sélectionner plus souvent le modèle à risques instantanés additifs quand le modèle de génération se faisait selon le modèle de Cox-Aalen de type 1 et le modèle à risques instantanés multiplicatifs quand le modèle de génération se faisait selon le modèle de Cox-Aalen de type 2. La comparaison entre les deux modèles à risques instantanés additifs et les deux modèles à risques instantanés multiplicatifs montrait

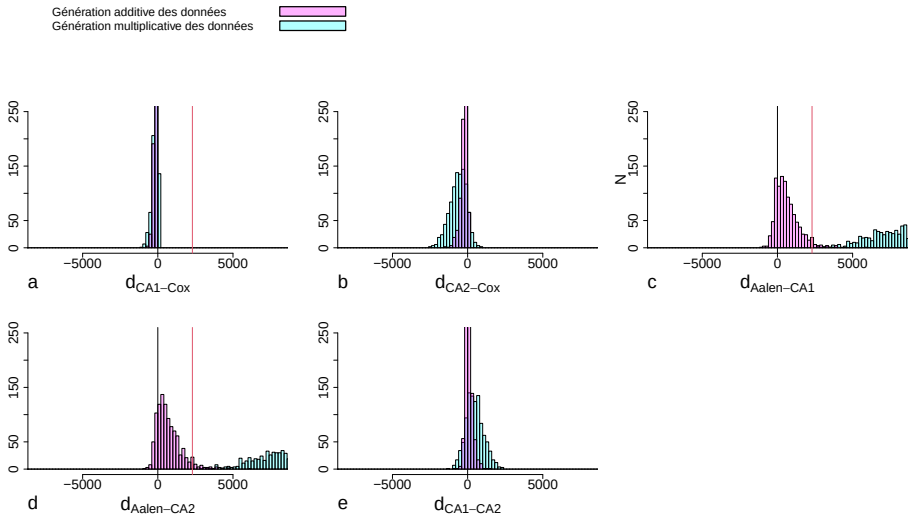


FIGURE 3.5. – Distributions des différences des SCPR estimées obtenues avec les modèles de Cox-Aalen, de Cox et d'Aalen comparés deux à deux. Situation avec deux covariables générées constantes et moyennes, une binaire et une continue, et un échantillon de 2 000 sujets.

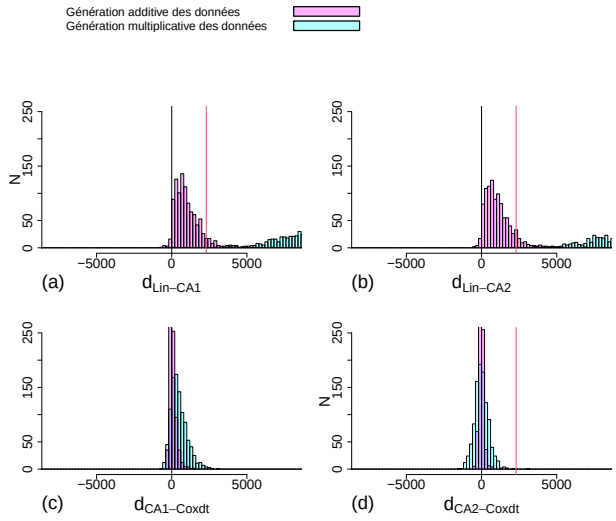


FIGURE 3.6. – Distributions des différences des SCPR estimées obtenues avec les modèles de Cox-Aalen, de Coxdt et de Lin comparés deux à deux. Situation avec deux covariables générées constantes et moyennes, une binaire et une continue, et un échantillon de 2 000 sujets.

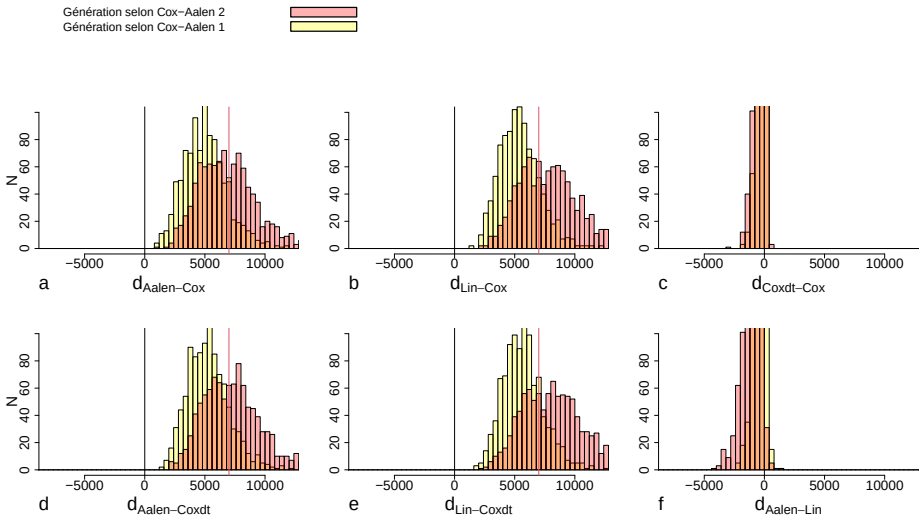


FIGURE 3.7. – Distributions des différences des SCPR estimées obtenues avec les 4 modèles à risques instantanés multiplicatifs ou additifs comparés deux à deux. Situation avec deux variables binaire et continue générées constantes et moyennes, l'une ayant un effet additif, l'autre un effet multiplicatif, et un échantillon de 2 000 sujets.

que les différences de SCPR étaient presque toujours négatives quel que soit le modèle de génération sans qu'il soit possible de trouver un seuil permettant de les discriminer.

Les distributions des différences de SCPR estimées par le modèle d'Aalen et les modèles de Cox-Aalen de type 1 et 2 étaient semblables à celle obtenue avec les modèles d'Aalen et de Cox, que la génération soit additive ou multiplicative (figure 3.8). Les modèles de Cox-Aalen de type 1 et 2 s'ajustaient mieux que le modèle d'Aalen quel que soit le modèle de génération. Les différences de SCPR entre les modèles de Cox-Aalen de type 1 ou 2 et le modèle de Cox étaient également petites. Le modèle de Cox-Aalen de type 2 semblait toutefois mieux s'ajuster que le modèle de Cox quand le modèle de génération était le modèle de Cox-Aalen de type 2 que lorsque le modèle de génération était le modèle de Cox-Aalen de type 1. En revanche, la comparaison des deux modèles de Cox-Aalen montrait une faible différence des SCPR estimées quel que soit le modèle de génération : le modèle de Cox-Aalen 1 était retenu dans 37,5 % des cas quand le modèle de génération était le modèle de Cox-Aalen de type 1 et le modèle de Cox-Aalen 2 était retenu dans 82,5 % des cas quand le modèle de génération était le modèle de Cox-Aalen de type 2.

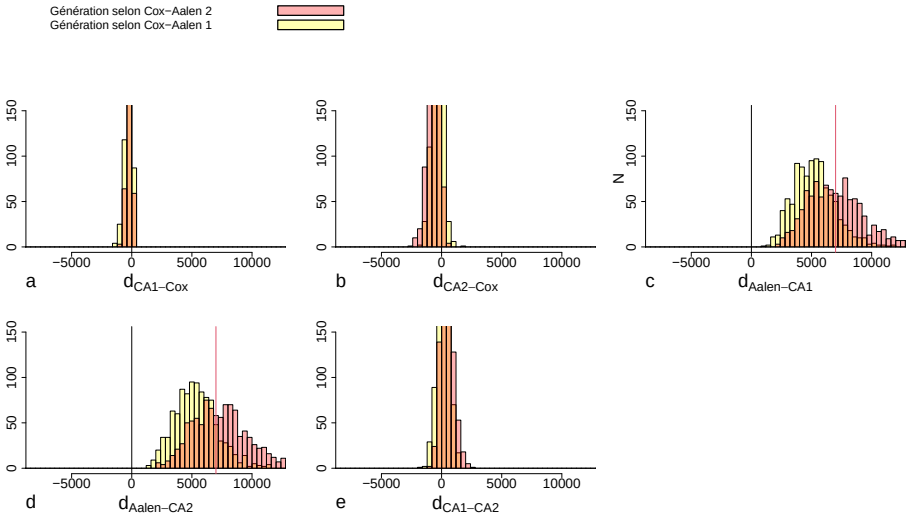


FIGURE 3.8. – Distributions des différences des SCPR estimées obtenues avec les modèles de Cox-Aalen, de Cox et d'Aalen comparés deux à deux. Situation avec deux variables binaire et continue générées constantes et moyennes, l'une ayant un effet additif, l'autre un effet multiplicatif, et un échantillon de 2 000 sujets.

Les distributions des différences de SCPR estimées par les modèles de Lin ou de Coxdt et les modèles de Cox-Aalen de type 1 et 2 étaient relativement proches de celles obtenues avec les modèles de Lin ou de Coxdt et le modèle de Cox, que la génération soit additive ou multiplicative (figure 3.9). De la même façon qu'avec le

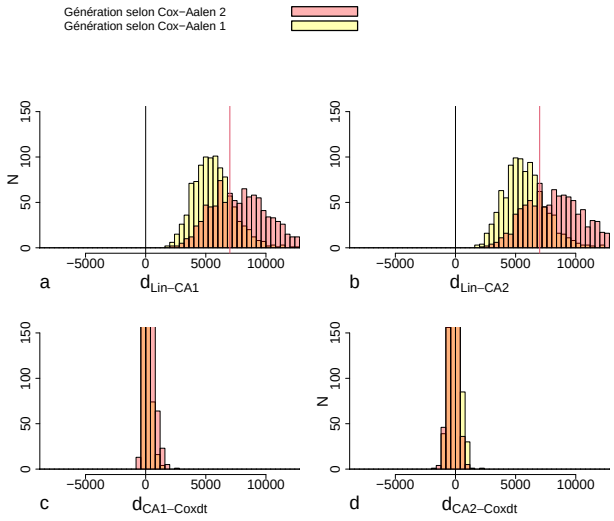


FIGURE 3.9. – Distributions des différences des SCPR estimées obtenues avec les modèles de Cox-Aalen, de Coxdt et de Lin comparés deux à deux. Situation avec deux variables binaire et continue générées constantes et moyennes, l'une ayant un effet additif, l'autre un effet multiplicatif, et un échantillon de 2 000 sujets.

modèle d'Aalen, les modèles de Cox-Aalen se comportaient comme des modèles de Cox en termes de pseudo-résidus.

La figure 3.10 permet de montrer la moyenne des SCPR calculées avec les six modèles selon le type de génération (multiplicative, additive, de type CA1 et de type CA2). Quand la génération était purement multiplicative ou simultanément multiplicative et additive, les modèles à risques instantanés additifs avaient des SCPR moyennes nettement plus élevées que les modèles à risques instantanés multiplicatifs ou simultanément multiplicatifs et additifs. En revanche, quand la modélisation était additive, les modèles à risques instantanés additifs avaient des SCPR moyennes très légèrement supérieures à celles les modèles à risques instantanés multiplicatifs ou simultanément multiplicatifs et additifs qui s'ajustaient donc également correctement aux données.

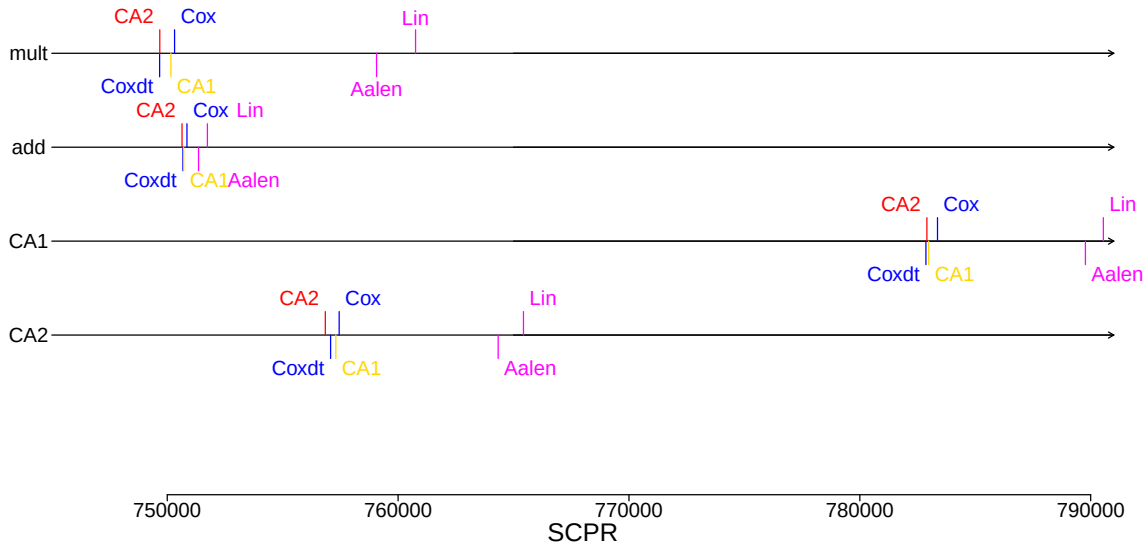


FIGURE 3.10. – Moyenne des pseudo-résidus calculés avec les six modèles en fonction de la génération des données. Situation avec deux covariables générées constantes et moyennes, une binaire et une continue, et un échantillon de 2 000 sujets.

Par conséquent, il est possible de proposer la règle de décision suivante :

— quand la SCPR obtenue par un modèle additif est inférieure à celles obtenues par les autres modèles, il convient de conserver le modèle à risques instantanés additif.
— quand la SCPR plus petite n'est pas obtenue par un modèle additif :
  — soit celles obtenues par les modèles additifs sont largement plus grandes auquel cas il faut garder le modèle avec la plus petite SCPR.
  — soit celles obtenues par les modèles additifs sont légèrement plus grandes auquel cas on peut retenir le modèle additif avec la plus petite SCPR.

3.4. Applications

Afin d'étudier l'intérêt de cette stratégie, celle-ci a été appliquée à trois exemples de données réelles.

3.4.1. Cirrhose biliaire primitive

Le premier exemple correspond à des données issues d'un essai randomisé de patients atteints d'une cirrhose biliaire primitive [18] dans lequel 418 patients ont été inclus de 1974 à 1984 et suivi jusqu'à leur décès ou leur censure. Ont été étudiés les effets de la bilirubine, une covariable continue, et la présence d'œdème, une covariable binaire, sur le risque instantané de décès. La stratégie de modélisation utilisée fut celle présentée au deuxième chapitre pour les deux types de modèle. Celle-ci a montré que pour une analyse avec un modèle multiplicatif, la forme fonctionnelle correcte de la bilirubine était logarithmique et que pour une analyse avec un modèle additif, la forme fonctionnelle correcte de la bilirubine était linéaire. La bilirubine a été modélisée avec les quatre modèles utilisés lors de l'étude de simulation (Cox, Coxdt, Lin et Aalen), et les SCPR ont été calculées (tableau 3.5).

Modèle	Covariable		
	bilirubine	ln(bilirubine)	œdème
modèle de Cox	44 060,91	41 335,56	47 658,22
modèle de Coxdt	43 718,12	41 346,08	47 529,79
modèle d'Aalen	41 434,94	43 639,39	47 488,67
modèle de Lin	41 795,10	43 246,28	47 624,72

Tableau 3.5. – SCPR estimées obtenus avec les quatre modèles et une seule covariable (bilirubine ou œedème) chez les patients ayant une cirrhose biliaire primitive.

Quand la bilirubine était modélisée sans transformation, les SCPR estimées obtenues avec les deux modèles à risques instantanés additifs étaient inférieures à celles obtenues avec les deux modèles multiplicatifs indiquant un meilleur ajustement avec un modèle additif. Par ailleurs, les comparaisons entre les modèles à risques instantanés multiplicatifs et entre les modèles à risques instantanés additifs montraient que les modèles plus souples (Coxdt et Aalen) avaient des SCPR plus petites que les modèles plus simples. Quand la bilirubine était modélisée avec une transformation logarithmique, les SCPR estimées obtenues avec les deux modèles à risques instantanés additifs étaient nettement supérieures à celles obtenues avec les deux modèles multiplicatifs indiquant un meilleur ajustement avec un modèle multiplicatif. Le modèle qui présentait la plus petite SCPR quelle que soit la modélisation de la bilirubine

était le modèle de Cox modélisé avec le logarithme de la bilirubine. Les analyses portant uniquement sur l'effet de l'œdème montrait une faible différence de SCPR entre les quatre modèles avec une SCPR inférieure pour le modèle d'Aalen confortant les résultats trouvés dans l'étude de simulations.

La modélisation simultanée des deux covariables, avec et sans transformation logarithmique de la bilirubine, a donné des résultats similaires à ceux obtenus avec la bilirubine seule (tableau 3.6). Concernant les modèles de Cox-Aalen, celui qui modélisait la bilirubine de manière multiplicative avait une SCPR proche des SCPR des modèles multiplicatifs et celui qui modélisait la bilirubine de manière additive avait une SCPR proche des SCPR des modèles additifs. C'est donc l'effet de la covariable continue qui semblait guider le comportement des modèles de Cox-Aalen en termes de SCPR.

Modèle	Covariable	
	bilirubine + œdème	ln(bilirubine) + œdème
modèle de Cox	42 999,50	40 195,68
modèle de Coxdt	42 470,91	40 179,02
modèle d'Aalen	40 305,74	42 772,02
modèle de Lin	40 717,10	42 522,07
modèle CA1 (bili mult, œdème add)	43 001,18	40 202,31
modèle CA2 (bili add, œdème mult)	40 509,13	41 944,75

Tableau 3.6. – SCPR estimées obtenues avec les six modèles et les deux covariable (bilirubine et œedème) chez les patients ayant une cirrhose biliaire primitive.

Le modèle de Cox avec un effet dépendant du temps linéaire ajusté sur l'œdème et le logarithme de la bilirubine avait la SCPR la plus faible et s'ajustait le mieux aux données.

3.4.2. Cancer du sein

Le second exemple correspond à des données de survie issues d'un essai clinique incluant 686 patientes atteintes d'un cancer du sein [46]. Les analyses portaient sur l'effet des œstrogènes, une covariable continue, et du traitement, une covariable binaire sur le risque de décès. La stratégie de modélisation suivie était celle décrite dans le deuxième chapitre et a permis de montrer que pour une analyse avec un modèle multiplicatif, la forme fonctionnelle correcte des œstrogènes était quadratique et que pour une analyse avec un modèle additif, la forme fonctionnelle correcte des œstrogènes était un spline quadratique avec trois nœuds correspondant à ses quartiles.

Modèle	Covariables			
	œst + œst^2	S(œst)	œst + œst^2 + trt	S(œst) + trt
modèle de Cox	99 203,27	98 003,28	98 434,95	97 330,53
modèle de Coxdt	98 995,38	97 805,12	98 233,96	97 136,90
modèle d'Aalen	99 865,54	97 884,79	99 657,04	97 062,02
modèle de Lin	99 232,61	98 029,51	98 526,79	97 216,38
modèle CA1 (œst mult, trait add)			98 410,24	97 315,62
modèle CA2 (œst add, trait mult)			98 672,00	97 184,20

Tableau 3.7. – SCPR estimées obtenues avec les six modèles et les deux covariable (œstrogènes et traitement) chez les patientes ayant un cancer du sein.

Quand les œstrogènes étaient modélisés avec une fonction quadratique, les modèles à risques instantanés multiplicatifs avaient des SCPR légèrement plus petites que les modèles à risques instantanés additifs. Quand les œstrogènes étaient modélisés avec une fonction spline quadratique, le modèle de Cox avec un effet dépendant du temps linéaire gardait la SCPR la plus basse (tableau 3.7) mais la SCPR obtenue avec le modèle d'Aalen devenait plus basse que celle obtenue avec le modèle de Cox. La modélisation avec les deux covariables montrait que le modèle d'Aalen ajusté sur les œstrogènes modélisés avec une fonction spline quadratique ajustait le mieux les données.

3.4.3. Infarctus du myocarde

Les pseudo-résidus ont également été calculés avec les modèles obtenus dans le deuxième chapitre. La SCPR du modèle à risques instantanés multiplicatifs était de 644 184,2 alors que celle du modèle à risques instantanés additifs était de 652 230,4. La grande différence entre les SCPR conduit à penser que le modèle à risques instantanés multiplicatifs s'ajustait mieux que celui à risques instantanés additifs.

3.5. Discussion

Dans ce chapitre, un nouvel outil diagnostique, les pseudo-résidus, a été proposé en analyse de survie et la SCPR a été utilisée comme méthode de sélection entre un modèle à risques instantanés multiplicatifs, un modèle à risques instantanés additifs et un modèle à risques instantanés simultanément multiplicatifs et additifs. Les pseudo-résidus peuvent aisément se calculer pour tout type de modèle car ils ne nécessitent que l'estimation des pseudo-observations et de la survie individuelle estimée à partir des différents modèles. Le modèle ayant la plus petite SCPR s'ajustant le mieux aux données, c'est donc celui-ci qu'il faut retenir. L'avantage principal de cette approche est qu'elle peut être généralisée à tout type de modèle à risques instantanés, qu'il soit

de nature additive, multiplicative ou simulaténement multiplicative et additive ainsi qu'aux autres modèles comme par exemple ceux à survie accélérée. De plus, cet outil peut être utilisé quel que soit le nombre de covariables incluses dans le modèle.

L'étude de simulations a montré que la somme des SCPR pouvait permettre de retenir le modèle approprié quand le modèle de génération incluait une seule variable continue. Quand le modèle de génération incluait une seule variable binaire, c'était le modèle d'Aalen qui était pratiquement toujours sélectionné. Ce résultat, à première vue surprenant, peut s'expliquer par le fait que les estimations du modèle d'Aalen dans ce cas sont identiques à celles de l'estimateur de Nelson-Aalen. De plus, la survie obtenue avec cet estimateur est très proche de celle que l'on obtient avec l'estimateur de Kaplan-Meier [13] qui est à la base du calcul des pseudo-observations et donc des pseudo-résidus. En absence de censures, on peut montrer (appendice 2) que la SCPR est égale à la somme des distances au carré entre l'estimation de Kaplan-Meier et celle obtenue avec le modèle d'Aalen. Par conséquent, la SCPR estimée avec le modèle d'Aalen est très souvent la plus petite. Ce dernier s'ajuste donc le mieux en présence d'une seule variable binaire. Toutefois, quand la proportion de censures ou la taille de l'échantillon augmente, il peut arriver, quand la génération des données est multiplicative, que le modèle de Cox ait une SCPR plus petite. De plus, en analyse de survie, excepté le cas des essais cliniques randomisés pour lesquels il n'est pas nécessaire d'ajuster sur d'autres covariables, il est possible d'utiliser un test. Dans le cadre des essais cliniques randomisés, il peut être plus pertinent d'utiliser un modèle d'Aalen également pour quantifier le gain dû au traitement en termes de risques instantanés cumulés. En présence d'une variable binaire, dont l'effet dépend du temps, générée de manière additive, la proportion de cas pour lesquels le modèle de Lin était retenu par rapport au modèle de Cox décroissait quand la génération était additive. Ceci s'explique par le fait qu'une variable binaire dont l'effet dépend du temps, générée de manière additive, se rapproche plus d'un effet multiplicatif constant que d'un effet additif constant.
Concernant les résultats obtenus avec deux covariables, la simple comparaison des pseudo-résidus ne permet pas à elle seule de retenir celui correspondant au modèle de génération. En effet, la différence des SCPR entre un modèle à risques instantanés additifs et un modèle à risques instantanés multiplicatifs est presque toujours négative à l'exception d'environ 25 % des cas uniquement quand la génération est additive. Ainsi, une différence négative permet de retenir avec une grande probabilité le modèle à risques instantanés additifs. Quand cette différence est positive, le choix peut être guidé par l'écart entre les SCPR obtenues avec les modèles à risques instantanés multiplicatifs et additifs.

L'approche appliquée aux données réelles de trois exemples a permis de montrer son intérêt dans le choix du meilleur modèle. Concernant les données de la cirrhose biliaire primitive, les résultats concernant l'analyse univariée confirment les connaissances que l'on avait obtenues grâce aux outils présentés au deuxième chapitre sur la forme fonctionnelle de la bilirubine. L'ajout de la covariable traitement montrait que le modèle de Coxdt avait non seulement la SCPR la plus faible mais aussi qu'elle était

assez fortement plus petite que celle obtenue avec les modèles à risques instantanés additifs confirmant le choix de le conserver. Pour les données sur le cancer du sein, le modèle avec la plus petite SCPR était le modèle d'Aalen avec une modélisation des œstrogènes avec un spline cubique à trois nœuds. C'est donc celui qui s'ajustait le mieux aux données. Enfin, pour les modèles obtenus dans le deuxième chapitre, les pseudo-résidus ont été calculés et ont montré que le modèle à risques instantanés multiplicatifs s'ajustait le mieux aux données bien qu'il fût légèrement plus complexe car possédant neuf paramètres, n'excluant pas une possible sur-paramétrisation.

Pour sélectionner un modèle à risques instantanés, différentes approches ont déjà été proposées et étudiées. Gandy et Jensen [19] ont proposé un test d'ajustement basé sur un pont brownien pour le modèle d'Aalen. Celui-ci repose sur la comparaison entre le modèle d'Aalen et différentes alternatives dont le modèle de Cox. Ce test est intéressant, toutefois, il repose sur l'hypothèse nulle selon laquelle le modèle d'Aalen s'ajuste aussi bien que le modèle de Cox. Aussi, il permet de rejeter le modèle d'Aalen quand il s'ajuste moins bien que le modèle de Cox, mais pas de rejeter le modèle de Cox quand il s'ajuste moins bien que le modèle d'Aalen. Gandy et Jensen ont comparé leurs résultats à ceux obtenus par McKeague et Utikal [34] qui ont proposé une approche différente. Celle-ci repose sur deux tests, le premier comparant le modèle de Cox à un estimateur non-paramétrique basé sur la fonction de risque instantané cumulée conditionnelle proposée par Beran [10], le second comparant le modèle d'Aalen au même estimateur non-paramétrique. L'inconvénient de cette approche est qu'elle utilise deux tests dont les résultats peuvent être discordants, c'est-à-dire rejeter ou ne pas rejeter simultanément les deux modèles sans que l'on puisse savoir lequel est le plus approprié. De plus, ces tests nécessitent des effectifs de plus de 3000 sujets pour présenter de bonnes propriétés. Aussi, pour comparer directement le modèle de Cox et celui d'Aalen, Martinussen, Aalen et Scheike [29] ont développé l'approche proposée par Mizon et Richard [35] consistant à tester des modèles non emboités en remplaçant le processus de comptage par son compensateur estimé avec l'autre modèle. Cette séduisante approche pose également le problème d'une potentielle discordance des résultats des deux tests et donc d'une potentielle difficulté du choix du meilleur modèle à conserver. Notre approche quant à elle permet toujours de retenir un modèle. Par ailleurs, ces trois approches ne sont pas disponibles dans les différents logiciels et les auteurs n'ont pas mis à disposition leurs codes rendant leur utilisation en pratique délicate.

Il existe toutefois des limites à l'approche que nous avons proposée. Si celle-ci permet de sélectionner un modèle parmi plusieurs, elle ne prend pas en compte la complexité des modèles. Par conséquent, il est possible que le modèle retenu s'ajuste effectivement mieux que ses concurrents mais au prix d'une plus grande complexité et donc d'un risque de sur-paramétrisation. Il pourrait donc être opportun d'ajouter une pénalisation afin de prendre en compte cette complexité qui pourrait faire l'objet de travaux futurs.

Valorisation scientifique

Ce travail a été valorisé scientifiquement :

Publication scientifique :
François Lefebvre, Roch Giorgi (2020). « An approach based on pseudo-residuals as a diagnostic tool for selecting a multiplicative or an additive hazards regression model ». Statistics in Medicine, XX(X).(Soumis)

Communications orales :
dans des conférences internationales à comité scientifique de sélection :
« Pseudo-residuals for selecting a multiplicative or an additive hazards regression model ». Lefebvre François, Giorgi R. 40^{th} annual conference of International Society for Clinical Biostatistics, Leuven, Belgique

dans des séminaires internes :
« An approach based on the pseudo-residuals for selecting a multiplicative or an additive hazards regression model », Seminár International Society for Clinical Biostatistics CZ, Prague, République Tchèque, 22/11/2019

« Utilisation des pseudo-residus pour la modélisation des covariables en analyse de survie », Webinar QuanTIM, 21/06/2019

Conclusion

Les modèles de régression à risques instantanés sont très utilisés dans le domaine biomédical car ils permettent de mesurer l'effet simultané de plusieurs covariables sur le risque instantané de base. Le modèle à risques instantanés proposé par Cox [15] jouit d'une très grande notoriété et est de fait le plus largement utilisé dans la littérature biomédicale. Il présente en effet de nombreux avantages, à savoir notamment sa souplesse due à la forme non-paramétrique du risque instantané de base et à l'obtention de paramètres dont l'exponentielle s'interprète comme un risque relatif. Ce modèle suppose que les covariables agissent de manière multiplicative sur le risque instantané de base, que l'effet des covariables continues est log-linéaire et que l'effet des covariables est constant au cours du temps. De nombreuses extensions du modèle de Cox ont été développées afin de modéliser des covariables dont les effets ne sont pas constants au cours du temps. Une autre approche a été proposée par Aalen [3] dont le modèle suppose un effet additif des covariables sur le risque instantané de base. Les résultats de ce modèle non-paramétrique sont des fonctions mesurant l'effet des covariables au cours du temps sans qu'une autre hypothèse que la linéarité ne soit nécessaire. Quand les effets de toutes les covariables sont constants au cours du temps, un modèle plus simple a été proposé par Lin et Ying [26]. Enfin, des modèles combinant les deux effets ont également été proposés, comme le modèle de Cox-Aalen [43].

La motivation de ce travail de thèse vient de l'intérêt qu'ont les modèles à risques instantanés additifs à décrire les effets des covariables au cours du temps et à les exprimer sous forme de différence de risques instantanés, c'est-à-dire ne dépendant pas du risque instantané de base. Dans un premier temps, comme il n'existait pas de méthode définie de modélisation avec un modèle à risques instantanés multiplicatifs et avec un modèle à risques instantanés additifs, une stratégie permettant d'analyser des données de survie avec chacun des deux modèles a été proposée. Pour cela, les outils diagnostiques habituels, comme les résidus de Schoenfeld [44] et les résidus de martingale [47], ont été utilisés avec d'autres outils moins connus, à savoir les pseudo-observations [39], les graphiques d'Arjas [9] et les processus de résidus de martingale [4]. Les pseudo-observations ont été développées dans le cadre de l'analyse de la survie et permettent de disposer d'une représentation non-paramétrique de l'effet des covariables au cours du temps. Les graphiques d'Arjas permettent de vérifier l'adéquation du modèle aux données en représentant les événements estimés en fonction des événements observés et les processus de résidus de martingale permettent de suivre l'évolution de la différence entre les événements observés et estimés au cours du temps. Comme ces outils sont moins connus, ils n'étaient pas,

excepté pour les pseudo-observations, disponibles dans les logiciels de statistiques habituellement utilisés comme R. Il a donc été nécessaire de les implémenter afin de permettre à toute personne analysant des données de survie de réaliser facilement la stratégie proposée. Celle-ci a été appliquée sur des données réelles et a montré qu'elle permettait d'obtenir des modèles à risques instantanés multiplicatifs et additifs qui s'ajustaient correctement aux données et qui différaient en termes d'interprétation.

Dans un second temps, la question du choix du meilleur modèle entre un modèle de régression à risques instantanés multiplicatifs, additifs et simultanément multiplicatifs et additifs a été étudiée. En effet, une fois la stratégie proposée appliquée, les modèles à risques instantanés multiplicatifs et additifs sont obtenus sans qu'il soit possible de déterminer celui qui s'ajuste le mieux aux données. Pour cela, les pseudos-résidus obtenus à partir des pseudo-observations et qui ont été proposés par Pohar-Perme [39], ont été étudiés. Comme ils représentent une distance entre une estimation non-paramétrique et une estimation obtenue par un modèle, la minimisation de la somme des carrés des pseudos-résidus a été proposée comme critère de sélection. Une étude de simulations a été réalisée pour étudier les performances de ce critère selon différents scénarios. Des données de survie ont été générées pour une variable qualitative puis pour une variable quantitative avec un effet multiplicatif puis avec un effet additif avec différentes tailles d'échantillon et différentes forces des effets. Elles ont ensuite été analysées avec deux modèles à risques instantanés multiplicatifs (le modèle de Cox et le modèle de Cox avec un effet dépendant du temps linéaire) et deux modèles à risques instantanés additifs (le modèle de Aalen et le modèle de Lin). Quatre autres scénarios de génération de données de survie ont été étudiés avec simultanément une variable qualitative et une variable quantitative ayant chacune soit un effet multiplicatif soit un effet additif sur le risque instantané de base. Ces données de survie ont été analysées avec six modèles, les quatre utilisés précédemment et deux modèles de Cox-Aalen différents. Ces analyses ont montré que ce critère permettait de sélectionner le modèle avec lequel ont été générées les données dans la plupart des cas pour une seule variable quantitative. En revanche, pour une seule variable qualitative, le modèle de Aalen s'ajuste presque toujours mieux que le modèle de Cox quelle que soit la méthode de génération des données, ce que l'on a pu établir mathématiquement. Pour la sélection de modèles avec deux covariables, une qualitative et une quantitative, les analyses ont montré que le modèle à risques instantanés additifs devait être retenu quand sa somme des carrés des pseudos-résidus était plus petite ou légèrement supérieure à celle des modèles à risques instantanés multiplicatifs, et que le modèle à risques instantanés multiplicatifs devait être retenu quand sa somme des carrés des pseudos-résidus était nettement plus petite que celle des modèles à risques instantanés additifs. Toutefois, cette somme ne permet pas de distinguer un modèle à risques instantanés multiplicatifs avec un modèle de Cox-Aalen qui semblent s'ajuster de manière analogue aux données.

Une perspective de ce travail serait d'étudier plus précisément l'écart entre les SCPR des modèles à risques instantanés mutliplicatifs et additifs afin de pouvoir déterminer plus précisément et objectivement un seuil permettant de sélectionner l'un des deux

modèles. Une piste pourrait être d'explorer par un développement analytique les facteurs et leur taille d'effet influençant la SCPR et de confirmer par une nouvelle étude de simulations avec plus de deux covariables avec éventuellement une interaction ces résultats dans un cadre plus général. En effet, il est déjà connu que cette somme dépend du nombre de sujets et de la proportion de censures mais sans en connaître la taille de l'effet. Elle pourrait également dépendre d'autres facteurs comme le nombre de covariables. Une autre perspective serait de prendre en compte la complexité des modèles dans le calcul des SCPR afin, dans une logique de parsimonie, de pouvoir pénaliser le modèle le plus complexe. Pour conclure, si l'intérêt des modèles à risques instantanés multiplicatifs dans l'analyse de survie n'est plus à démontrer, l'intérêt des modèles à risques instantanés additifs n'est pas encore assez connu et doit être développé afin de permettre la réalisation du modèle le plus adapté dont le choix final peut être réalisé en utilisant la somme des carrés des pseudo-résidus.

Bibliographie

[1] Odd AALEN. « Nonparametric inference for a family of counting processes ». In : *The Annals of Statistics* (1978), p. 701-726 (cf. p. 13, 18).

[2] Odd AALEN, Ornulf BORGAN et Hakon GJESSING. *Survival and event history analysis : a process point of view*. Springer Science & Business Media, 2008 (cf. p. 54).

[3] Odd O AALEN. « A linear regression model for the analysis of life times ». In : *Statistics in medicine* 8.8 (1989), p. 907-925 (cf. p. 24, 26, 80).

[4] Odd O AALEN. « Further results on the non-parametric linear regression model in survival analysis ». In : *Statistics in medicine* 12.17 (1993), p. 1569-1588 (cf. p. 80).

[5] Michal ABRAHAMOWICZ et Todd A MACKENZIE. « Joint estimation of time-dependent and non-linear effects of continuous covariates on survival ». In : *Statistics in medicine* 26.2 (2007), p. 392-408 (cf. p. 30, 54).

[6] Per Kragh ANDERSEN et Richard D GILL. « Cox's regression model for counting processes : a large sample study ». In : *The annals of statistics* (1982), p. 1100-1120 (cf. p. 14, 20).

[7] Per Kragh ANDERSEN, John P KLEIN et Susanne ROSTHØJ. « Generalised linear models for correlated pseudo-observations, with applications to multi-state models ». In : *Biometrika* 90.1 (2003), p. 15-27 (cf. p. 31).

[8] Per Kragh ANDERSEN et Maja POHAR PERME. « Pseudo-observations in survival analysis ». In : *Statistical methods in medical research* 19.1 (2010), p. 71-99 (cf. p. 31).

[9] Elja ARJAS. « A graphical method for assessing goodness of fit in Cox's proportional hazards model ». In : *Journal of the American Statistical Association* 83.401 (1988), p. 204-212 (cf. p. 32, 80).

[10] Rudolf BERAN. « Nonparametric regression with randomly censored survival data ». In : (1981) (cf. p. 78).

[11] PE BÖHMER. « Theorie der unabhängigen Wahrscheinlichkeiten ». In : *Rapports Memoires et Proces verbaux de Septieme Congres International dActuaires Amsterdam*. T. 2. 1912, p. 327-343 (cf. p. 13).

[12] NE BRESLOW et NE DAY. « Statistical Methods in Cancer Research. Volume II–The Design and Analysis of Cohort Studies ». In : *IARC scientific publications* 82 (1987), p. 1-406 (cf. p. 54).

[13] Enrico COLOSIMO, Flavio FERREIRA, Maristela OLIVEIRA et al. « Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators ». In : *Journal of Statistical Computation and Simulation* 72.4 (2002), p. 299-308 (cf. p. 77).

[14] Giuliana CORTESE, Thomas H SCHEIKE et Torben MARTINUSSEN. « Flexible survival regression modelling ». In : *Statistical methods in medical research* 19.1 (2010), p. 5-28 (cf. p. 54).

[15] David R COX. « Regression models and life-tables ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 34.2 (1972), p. 187-202 (cf. p. 13, 19, 80).

[16] David R COX. « Partial likelihood ». In : *Biometrika* 62.2 (1975), p. 269-276 (cf. p. 20).

[17] Bradley EFRON. « The efficiency of Cox's likelihood function for censored data ». In : *Journal of the American statistical Association* 72.359 (1977), p. 557-565 (cf. p. 20).

[18] Thomas R FLEMING et David P HARRINGTON. *Counting processes and survival analysis*. T. 169. John Wiley & Sons, 2011 (cf. p. 74).

[19] Axel GANDY et UWE JENSEN. « On goodness-of-fit tests for Aalen's additive risk model ». In : *Scandinavian Journal of Statistics* 32.3 (2005), p. 425-445 (cf. p. 78).

[20] Major GREENWOOD et al. « A report on the natural duration of cancer. » In : *A Report on the Natural Duration of Cancer.* 33 (1926) (cf. p. 13, 18).

[21] Kenneth R HESS. « Graphical methods for assessing violations of the proportional hazards assumption in Cox regression ». In : *Statistics in medicine* 14.15 (1995), p. 1707-1723 (cf. p. 22).

[22] Fred W HUFFER et Ian W MCKEAGUE. « Weighted least squares estimation for Aalen's additive risk model ». In : *Journal of the American Statistical Association* 86.413 (1991), p. 114-129 (cf. p. 25).

[23] Edward L KAPLAN et Paul MEIER. « Nonparametric estimation from incomplete observations ». In : *Journal of the American statistical association* 53.282 (1958), p. 457-481 (cf. p. 13).

[24] R KAY. « Goodness of fit methods for the proportional hazards regression model : a review. » In : *Revue d'épidémiologie et de santé publique* 32.3-4 (1984), p. 185 (cf. p. 22).

[25] John P KLEIN et Per Kragh ANDERSEN. « Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function ». In : *Biometrics* 61.1 (2005), p. 223-229 (cf. p. 31).

[26] DY LIN et Zhiliang YING. « Semiparametric analysis of the additive risk model ». In : *Biometrika* 81.1 (1994), p. 61-71 (cf. p. 27, 80).

[27] DY LIN, Zhiliang YING et al. « Semiparametric analysis of general additive-multiplicative hazard models for counting processes ». In : *The annals of Statistics* 23.5 (1995), p. 1712-1734 (cf. p. 28).

[28] Chengyuan LU, Jelle GOEMAN et Hein PUTTER. « Maximum likelihood estimation in the additive hazards model ». In : *arXiv preprint arXiv :2004.06156* (2020) (cf. p. 57).

[29] Torben MARTINUSSEN, Odd O AALEN et Thomas H SCHEIKE. « The Mizon–Richard Encompassing Test for the Cox and Aalen Additive Hazards Models ». In : *Biometrics* 64.1 (2008), p. 164-171 (cf. p. 78).

[30] Torben MARTINUSSEN et Thomas H SCHEIKE. « A semiparametric additive regression model for longitudinal data ». In : *Biometrika* 86.3 (1999), p. 691-702 (cf. p. 55).

[31] Torben MARTINUSSEN et Thomas H SCHEIKE. « A flexible additive multiplicative hazard model ». In : *Biometrika* 89.2 (2002), p. 283-298 (cf. p. 29).

[32] Torben MARTINUSSEN et Thomas H SCHEIKE. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007 (cf. p. 55).

[33] Ian W MCKEAGUE et Peter D SASIENI. « A partly parametric additive risk model ». In : *Biometrika* 81.3 (1994), p. 501-514 (cf. p. 27).

[34] Ian W MCKEAGUE et Klaus J UTIKAL. « Goodness-of-fit tests for additive hazards and proportional hazards models ». In : *Scandinavian Journal of Statistics* (1991), p. 177-195 (cf. p. 55, 78).

[35] Grayham E MIZON et Jean-Francois RICHARD. « The encompassing principle and its application to testing non-nested hypotheses ». In : *Econometrica : Journal of the Econometric Society* (1986), p. 657-678 (cf. p. 78).

[36] Govind S MUDHOLKAR et Deo Kumar SRIVASTAVA. « Exponentiated Weibull family for analyzing bathtub failure-rate data ». In : *IEEE transactions on reliability* 42.2 (1993), p. 299-302 (cf. p. 59).

[37] Wayne NELSON. « Hazard plotting for incomplete failure data ». In : *Journal of Quality Technology* 1.1 (1969), p. 27-52 (cf. p. 13, 18).

[38] Wayne NELSON. « Theory and applications of hazard plotting for censored failure data ». In : *Technometrics* 14.4 (1972), p. 945-966 (cf. p. 13, 18).

[39] Maja Pohar PERME et Per Kragh ANDERSEN. « Checking hazard regression models using pseudo-observations ». In : *Statistics in Medicine* 27.25 (2008), p. 5309-5328 (cf. p. 58, 80, 81).

[40] Richard PETO et Julian PETO. « Asymptotically efficient rank invariant test procedures ». In : *Journal of the Royal Statistical Society : Series A (General)* 135.2 (1972), p. 185-198 (cf. p. 13).

[41] Maurice H QUENOUILLE. « Notes on bias in estimation ». In : *Biometrika* 43.3/4 (1956), p. 353-360 (cf. p. 31).

[42] Peter D SASIENI et Angela WINNETT. « Martingale difference residuals as a diagnostic tool for the Cox model ». In : *Biometrika* 90.4 (2003), p. 899-912 (cf. p. 30, 54).

[43] Thomas H SCHEIKE et Mei-jie ZHANG. « An additive–multiplicative Cox–Aalen regression model ». In : *Scandinavian Journal of Statistics* 29.1 (2002), p. 75-88 (cf. p. 28, 54, 80).

[44] David SCHOENFELD. « Partial residuals for the proportional hazards regression model ». In : *Biometrika* 69.1 (1982), p. 239-241 (cf. p. 22, 80).

[45] Judith D SINGER, John B WILLETT, John B WILLETT et al. *Applied longitudinal data analysis : Modeling change and event occurrence*. Oxford university press, 2003 (cf. p. 23).

[46] Terry M THERNEAU et Patricia M GRAMBSCH. « The Cox model ». In : *Modeling survival data : extending the Cox model*. Springer, 2000, p. 39-77 (cf. p. 22, 75).

[47] Terry M THERNEAU, Patricia M GRAMBSCH et Thomas R FLEMING. « Martingale-based residuals for survival models ». In : *Biometrika* 77.1 (1990), p. 147-160 (cf. p. 23, 80).

[48] Xianhong XIE, Howard D STRICKLER et Xiaonan XUE. « Additive hazard regression models : an application to the natural history of human papillomavirus ». In : *Computational and Mathematical Methods in Medicine* 2013 (2013) (cf. p. 26).

ANNEXES

A. Relation entre la somme des carrés des pseudo-residus et l'estimation de Kaplan-Meier

Pour un sujet i au temps t_s, le pseudo-résidu s'écrit :

$$\hat{\epsilon}_i(t_s) = n\hat{S}(t_s) - (n-1)\hat{S}^{-i}(t_s) - \hat{S}(t_s|x_i)$$

et la somme des carrés des pseudo-résidus pour l'ensemble des sujets au temps t_s est :

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2(t_s) = \sum_{i=1}^{n} \left(n\hat{S}(t_s) - (n-1)\hat{S}^{-i}(t_s) - \hat{S}(t_s|x_i) \right)^2$$

Soient X une variable binaire et $\hat{S}(t|x_i = 0) = \hat{S}_0(t)$ et $\hat{S}(t|x_i = 1) = \hat{S}_1(t)$ la survie estimée avec un modèle à risques instantanés pour les sujets avec $x = 0$ et $x = 1$, respectivement. En cas d'absence de censures et de temps de survie identiques, au temps t_s, s sujets ont présenté l'événement d'intérêt (s_0 pour les sujets avec $x = 0$ et s_1 pour les sujets avec $x = 1$). Au temps t_s, si un sujet avec $x = 0$ a présenté l'événement, la somme des carrés des pseudo-résidus est :

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2(t_s) = \sum_{i=1}^{s_0} \left(n\hat{S}(t_s) - (n-1)\hat{S}^{-i}(t_s) - \hat{S}_0(t_s) \right)^2 + \sum_{i=s_0+1}^{n_0} \left(n\hat{S}(t_s) - (n-1)\hat{S}^{-i}(t_s) - \hat{S}_0(t_s) \right)^2$$

$$+ \sum_{i=n_0+1}^{n_0+s_1} \left(n\hat{S}(t_s) - (n-1)\hat{S}^{-i}(t_s) - \hat{S}_1(t_s) \right)^2 + \sum_{i=n_0+s_1+1}^{n} \left(n\hat{S}(t_s) - (n-1)\hat{S}^{-i}(t_s) - \hat{S}_1(t_s) \right)^2$$

Pour tous les sujets, on a : $\hat{S}(t_s) = \frac{n-s}{n}$;
Pour les sujets qui ont présenté l'événement avant ou au temps t_s, $\hat{S}^{-i}(t_s) = \frac{n-s}{n-1}$;
Pour les sujets qui ont présenté l'événement après le temps t_s, $\hat{S}^{-i}(t_s) = \frac{n-s-1}{n-1}$.
Par conséquent :

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2(t_s) = \sum_{i=1}^{s_0} \left(n - s - (n-s) - \hat{S}_0(t_s) \right)^2 + \sum_{i=s_0+1}^{n_0} \left(n - s - (n-s-1) - \hat{S}_0(t_s) \right)^2$$

$$+ \sum_{i=n_0+1}^{n_0+s_1} \left(n - s - (n-s) - \hat{S}_1(t_s) \right)^2 + \sum_{i=n_0+s_1+1}^{n} \left(n - s - (n-s-1) - \hat{S}_1(t_s) \right)^2$$

En utilisant l'estimation de Kaplan-Meier $\hat{S}_{KM0}(t_s)$ et $\hat{S}_{KM1}(t_s)$ pour les sujets avec $x = 0$ et $x = 1$ respectivement, on a :

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2(t_s) = \sum_{i=1}^{s_0} \left(-\hat{S}_{KM0}(t_s) + \hat{S}_{KM0}(t_s) - \hat{S}_0(t_s) \right)^2 + \sum_{i=s_0+1}^{n_0} \left(1 - \hat{S}_{KM0}(t_s) + \hat{S}_{KM0}(t_s) - \hat{S}_0(t_s) \right)^2$$

$$+\sum_{i=n_0+1}^{n_0+s_1}\left(-\hat{S}_{KM1}(t_s)+\hat{S}_{KM1}(t_s)-\hat{S}_1(t_s)\right)^2+\sum_{i=n_0+s_1+1}^{n}\left(1-\hat{S}_{KM1}(t_s)+\hat{S}_{KM1}(t_s)-\hat{S}_1(t_s)\right)^2$$

Les deux premiers termes sont :

$$\sum_{i=1}^{s_0}\left(-\hat{S}_{KM0}(t_s)+\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2+\sum_{i=s_0+1}^{n_0}\left(1-\hat{S}_{KM0}(t_s)+\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2$$

$$=\sum_{i=1}^{s_0}\left[\left(\hat{S}_{KM0}(t_s)\right)^2-2\hat{S}_{KM0}(t_s)\left(-\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)+\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2\right]$$

$$+\sum_{i=s_0+1}^{n_0}\left[\left(1-\hat{S}_{KM0}(t_s)\right)^2-2\left(1-\hat{S}_{KM0}(t_s)\right)\left(-\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)+\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2\right]$$

$$=s_0\left(\hat{S}_{KM0}(t_s)\right)^2+(n_0-s_0)\left(1-\hat{S}_{KM0}(t_s)\right)^2+2\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)\left(-s_0\hat{S}_{KM0}(t_s)+(n_0-s_0)\left(1-\hat{S}_{KM0}(t_s)\right)\right)^2$$
$$+n_0\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2$$

$$=s_0\left(\hat{S}_{KM0}(t_s)\right)^2+(n_0-s_0)\left(1-\hat{S}_{KM0}(t_s)\right)^2+2\left(-\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)\left(-s_0\frac{(n_0-s_0)}{n_0}\right.$$
$$\left.+(n_0-s_0)\frac{s_0}{n_0}\right)+n_0\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2$$
$$=s_0\left(\hat{S}_{KM0}(t_s)\right)^2+(n_0-s_0)\left(1-\hat{S}_{KM0}(t_s)\right)^2+n_0\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2$$

Les deux derniers termes sont :

$$\sum_{i=n_0+1}^{n_0+s_1}\left(-\hat{S}_{KM1}(t_s)+\hat{S}_{KM1}(t_s)-\hat{S}_1(t_s)\right)^2+\sum_{i=n_0+s_1+1}^{n}\left(1-\hat{S}_{KM1}(t_s)+\hat{S}_{KM1}(t_s)-\hat{S}_1(t_s)\right)^2$$

$$=s_1\left(\hat{S}_{KM1}(t_s)\right)^2+(n_1-s_1)\left(1-\hat{S}_{KM1}(t_s)\right)^2+n_1\left(\hat{S}_{KM1}(t_s)-\hat{S}_1(t_s)\right)^2$$

On obtient finalement :

$$\sum_{i=1}^{n}\hat{\epsilon}_i^2(t_s)=s_0\left(\hat{S}_{KM0}(t_s)\right)^2+(n_0-s_0)\left(1-\hat{S}_{KM0}(t_s)\right)^2+s_1\left(\hat{S}_{KM1}(t_s)\right)^2+(n_1-s_1)\left(1-\hat{S}_{KM1}(t_s)\right)^2+$$
$$n_0\left(\hat{S}_{KM0}(t_s)-\hat{S}_0(t_s)\right)^2+n_1\left(\hat{S}_{KM1}(t_s)-\hat{S}_1(t_s)\right)^2$$

La somme des carrés des pseudo-résidus pour l'ensemble des n sujets au temps t_s est donc égale à la somme d'un terme constant et du carré des distances entre une estimation de Kaplan-Meier et une estimation obtenue pour chacun des deux

groupes.

B. Programme R

```
library(survival)
library(timereg)
library(ahaz)
data(TRACE)

# Addition d'un nombre aléatoire afin d'obtenir des temps de survie différents
pour chaque sujet
set.seed(906)
TRACE$time=TRACE$time+rnorm(nrow(TRACE),0,0.0001)

# Transformation des différentes causes de décès en une variable binaire
(décès/censure)
TRACE$status2=as.numeric(TRACE$status!=0)

# Classement des durée de survie de la plus petite à la plus grande
B=TRACE[order(TRACE[,"time"]),]
names(B)=c("id","wmi","status","chf","age","sex","dia","time","vf","status2")

# Proposition d'une stratégie de modélisation des modèles à risques instantanés
additifs
# Hypothèse de log-linéarité

# modèle de Cox vide
mCox0=coxph(Surv(time,status!=0)~1,data=B)

# résidus de martingale avec un modèle de Cox vide
MartResiduals0=residuals(mCox0,type="martingale")

# modèle de Cox avec exp(age/100)
mCoxe=coxph(Surv(time,status!=0)~I(exp(age/100)),data=B)

# résidus de martingale avec un modèle de Cox avec exp(age/100)
MartResidualse=residuals(mCoxe,type="martingale")

# Figure 1
x11();par(mfrow=c(1,2))
plot(B$age,MartResiduals0,xlab="âge (ans)",ylab="Résidus de martingale ",
cex.axis=1.7,cex.lab=1.7,cex.main=2)
abline(h=0,lty=2)
lines(lowess(B$age,MartResiduals0,iter=0))
mtext("a",side=1,adj=0,line=3.5,cex=2.5)
```

```
plot(B$age,MartResidualse,xlab="âge (ans)",ylab="Résidus de martingales"
,cex.axis=1.7,cex.lab=1.7,cex.main=2)
abline(h=0,lty=2)
lines(lowess(B$age,MartResidualse,iter=0))
mtext("b",side=1,adj=0,line=3.5,cex=2.5)

# Figure A1
# modèle de Cox avec un effet quadratique
mCox2=coxph(Surv(time,status!=0)~age+I(age^2),data=B)
MartResiduals2=residuals(mCox2,type="martingale")

# résidus de martingale avec un modèle de Cox avec un effet quadratique
MartResiduals2=residuals(mCox2,type="martingale")

x11();par(mfrow=c(1,2))
plot(B$age,MartResiduals0,xlab="âge",ylab="Résidus de martingales",
main="modèle de Cox vide",cex.axis=1.7,cex.lab=1.7,cex.main=2)
abline(h=0,lty=2)
lines(lowess(B$age,MartResiduals0,iter=0))
mtext("a",side=1,adj=0,line=3.5,cex=1.8)

plot(B$age,MartResiduals2,xlab="age",ylab="Résidus de martingales",
main="modèle de Cox avec un effet quadratique de l'âge ",cex.axis=1.7,
cex.lab=1.7,cex.main=2)
abline(h=0,lty=2)
lines(lowess(B$age,MartResiduals2,iter=0))
mtext("b",side=1,adj=0,line=3.5,cex=1.8)

# Comparaison des AIC
AIC(mCoxe)
AIC(mCox2)

# Tests de la corrélation des résidus de Schoenfeld
mCoxagee=coxph(Surv(time,status!=0)~I(exp(age/100)),data=B)
testagee<-cox.zph(mCoxagee,global=T,transform="rank")
print(testagee)

mCoxsex=coxph(Surv(time,status!=0)~sex,data=B)
testsex<-cox.zph(mCoxsex,global=T,transform="rank")
print(testsex)

mCoxchf=coxph(Surv(time,status!=0)~chf,data=B)
```

```
testchf<-cox.zph(mCoxchf,global=T,transform="rank")
print(testchf)

mCoxdia=coxph(Surv(time,status!=0)~dia,data=B)
testdia<-cox.zph(mCoxdia,global=T,transform="rank")
print(testdia)

mCoxvf=coxph(Surv(time,status!=0)~vf,data=B)
testvf<-cox.zph(mCoxvf,global=T,transform="rank")
print(testvf)


# Figure 2
x11();par(mfrow=c(2,3))
par(cex.axis=2)
plot(testagee,xlab="",ylab="",main="exp(age/100)",cex.lab=1.7,cex.main=3,lwd=3)
abline(h=mCoxagee$coefficients,lty=3,lwd=2)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)
mtext("Time (years)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Beta(t)",side=2,adj=0.5,line=2.2,cex=1.7)

plot(testsex,xlab="",ylab="",main="sex",cex.lab=1.7,cex.main=3,lwd=3)
abline(h=mCoxsex$coefficients,lty=3,lwd=2)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)
mtext("Time (years)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Beta(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(testchf,xlab="",ylab="",main="chf",cex.lab=1.7,cex.main=3,lwd=3)
abline(h=mCoxchf$coefficients,lty=3,lwd=2)
mtext("c",side=1,adj=0,line=3.5,cex=2.5)
mtext("Time (years)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Beta(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(testdia,xlab="",ylab="",main="dia",cex.lab=1.7,cex.main=3,lwd=3)
abline(h=mCoxdia$coefficients,lty=3,lwd=2)
mtext("d",side=1,adj=0,line=3.5,cex=2.5)
mtext("Time (years)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Beta(t)",side=2,adj=0.5,line=2.2,cex=1.7)

plot(testvf,xlab="",ylab="",main="vf",cex.lab=1.7,cex.main=3,lwd=3)
abline(h=mCoxvf$coefficients,lty=3,lwd=2)
mtext("e",side=1,adj=0,line=3.5,cex=2.5)
mtext("Time (years)",side=1,adj=0.5,line=3.5,cex=1.7)
```

```
mtext("Beta(t)",side=2,adj=0.5,line=2.5,cex=1.7)

# modèle avec un effet linéaire dépendant du temps pour l'ic
cut.points=unique(B$time[B$status2==1])
B2=survSplit(data=B,cut=cut.points,end="time",start="time0",event="status2")
B2$chft=B2$chf*B2$time
mCoxchf2=coxph(Surv(time0,time,status2)~chf+chft,data=B2)
summary(mCoxchf2)

testchf2<-cox.zph(mCoxchf2,global=T,transform="rank")
print(testchf2)

# modèle avec une rupture de pente pour la fv
# choix du seuil
z=NULL
cut.points=unique(B$time[B$status2==1])
B2=survSplit(data=B,cut=cut.points,end="time",start="time0",event="status2")
B2$vft=B2$vf*B2$time
for (i in 1:240){
  B2$vft2=B2$vf*(B2$time-i/100)*(B2$time>i/100)
  mCoxvf3=coxph(Surv(time0,time,status2)~vf+vft+vft2,data=B2)
  z[i]=AIC(mCoxvf3)
  cat(i,"\n");flush.console()
}

z2=NULL
for (i in 1:20){
  B2$vft2=B2$vf*(B2$time-(i+10)/100)*(B2$time>(i+10)/100)
  mCoxvf3=coxph(Surv(time0,time,status2)~vf+vft+vft2,data=B2)
  z2[i]=AIC(mCoxvf3)}

# AIC obtenu pour le seuil retenu
z[which(z==min(z))]

# seuil retenu
which(z==min(z))/100

# modèle de Cox avec le meilleur seuil
cut.points=unique(B$time[B$status2==1])
B2=survSplit(data=B,cut=cut.points,end="time",start="time0",event="status2")
B2$vft=B2$vf*B2$time
B2$vft2=B2$vf*(B2$time-0.15)*(B2$time>0.15)
mCoxvf3=coxph(Surv(time0,time,status2)~vf+vft+vft2,data=B2)
```

```
summary(mCoxvf3)
testmCoxvf3=cox.zph(mCoxvf3,global=T,transform="km")
print(testmCoxvf3)


# Qualité de l'ajustement
# Calcul des pseudo-observations
B$censure=as.numeric(B$status!=0)
KM0=matrix(0,nrow=nrow(B),ncol=nrow(B)-1)
for (i in 1:nrow(B)){B2=B[-i,]
for(j in 1:nrow(B2)){
  KM0[i,j]=1-B2$censure[j]/(nrow(B2)-j+1)}}
S0=matrix(0,nrow=nrow(B),ncol=nrow(B)-1)
for (i in 1:nrow(B)){S0[i,]=cumprod(KM0[i,])}
S=matrix(0,nrow=nrow(B),ncol=nrow(B))
S[1,2:nrow(B)]=S0[1,]
S[nrow(B),1:(nrow(B)-1)]=S0[nrow(B),]
S[nrow(B),nrow(B)]=S0[nrow(B),nrow(B)-1]
for (i in 2:(nrow(B)-1)){
  S[i,1:(i-1)]=S0[i,1:(i-1)]
  S[i,i]=S0[i,i-1]
  S[i,(i+1):(nrow(B))]=S0[i,i:(nrow(B)-1)]}
S[1,1]=1
KMt=rep(0,nrow(B))
for (i in 1:nrow(B)){KMt[i]=1-B$censure[i]/(nrow(B)-i+1)}
St=rep(0,nrow(B))
St=cumprod(KMt)

matS=matrix(rep(nrow(B)*St,nrow(B)),nrow=nrow(B),ncol=nrow(B),byrow=T)
POS=matS-(nrow(B)-1)*S

# Estimation aux neuf déciles des temps d'événement
times<-unique(B$time[B$status!=0])
tps=round(c(length(times)/10,2*length(times)/10,3*length(times)/10,
4*length(times)/10,5*length(times)/10,6*length(times)/10,
7*length(times)/10,8*length(times)/10,9*length(times)/10),0)

# Figure 3
x11()
plot(log(-log(lowess(B$age/100,POS[,tps[1]],iter=0)$y))~lowess(B$age/100,
POS[,tps[1]],iter=0)$x,xlab="age/100
(ans)",ylab="",type="l",ylim=c(-7.2,1.2),cex.axis=1.7,lwd=3,cex.lab=1.7,
col="grey0")
```

```
points(0.682305,-2.998537,pch=0,lwd=3,cex=1.5,col="grey0")
lines(log(-log(lowess(B$age/100,POS[,tps[2]],iter=0)$y))~lowess(B$age/100,
POS[,tps[2]],iter=0)$x,type="l",col="grey9",lwd=3)
points(0.682305,-2.337397,pch=1,lwd=3,cex=1.5,col="grey9")
lines(log(-log(lowess(B$age/100,POS[,tps[3]],iter=0)$y))~lowess(B$age/100,
POS[,tps[3]],iter=0)$x,type="l",col="grey18",lwd=3)
points(0.682305,-1.901599,pch=2,lwd=3,cex=1.5,col="grey18")
lines(log(-log(lowess(B$age/100,POS[,tps[4]],iter=0)$y))~lowess(B$age/100,
POS[,tps[4]],iter=0)$x,type="l",col="grey27",lwd=3)
points(0.682305,-1.586416,pch=3,lwd=3,cex=1.5,col="grey27")
lines(log(-log(lowess(B$age/100,POS[,tps[5]],iter=0)$y))~lowess(B$age/100,
POS[,tps[5]],iter=0)$x,type="l",col="grey36",lwd=3)
points(0.682305,-1.28576,pch=4,lwd=3,cex=1.5,col="grey36")
lines(log(-log(lowess(B$age/100,POS[,tps[6]],iter=0)$y))~lowess(B$age/100,
POS[,tps[6]],iter=0)$x,type="l",col="grey45",lwd=3)
points(0.682305,-1.046544,pch=5,lwd=3,cex=1.5,col="grey45")
lines(log(-log(lowess(B$age/100,POS[,tps[7]],iter=0)$y))~lowess(B$age/100,
POS[,tps[7]],iter=0)$x,type="l",col="grey54",lwd=3)
points(0.682305,-0.8717967,pch=6,lwd=3,cex=1.5,col="grey54")
lines(log(-log(lowess(B$age/100,POS[,tps[8]],iter=0)$y))~lowess(B$age/100,
POS[,tps[8]],iter=0)$x,type="l",col="grey63",lwd=3)
points(0.682305,-0.6818521,pch=7,lwd=3,cex=1.5,col="grey63")
lines(log(-log(lowess(B$age/100,POS[,tps[9]],iter=0)$y))~lowess(B$age/100,
POS[,tps[9]],iter=0)$x,type="l",col="grey72",lwd=3)
points(0.682305,-0.4755059,pch=8,lwd=3,cex=1.5,col="grey72")
abline(h=0)
legend("bottomright",c("1er décile (0,03 an)","2e décile (0,09 an)","3e décile
(0,37 an)","4e décile (0,97 an)","5e décile (1,73 an)","6e décile (2,46 ans)",
"7e décile (3,46 ans)","8e décile (4,61 ans)","9e décile (5,81
ans)"),col=c("grey0","grey9","grey18","grey27","grey36","grey45","grey54",
"grey63","grey72"),lwd=3,cex=1,pch=0:8,pt.cex=1.5)
mtext("cloglog(Pseudo-observations)",side=2,adj=0.5,line=2.5,cex=1.7)

# graphiques d'Arjas
# ic
mCoxchf=coxph(Surv(time,status!=0)~chf,data=B)

Lambdachf=matrix(0,nrow=nrow(B),ncol=nrow(B))
for (i in 1:nrow(B)){
  Lambdachf[i,1:i]=survfit(mCoxchf,newdata=data.frame(chf=B$chf[i]))$cumhaz[1:i]
  cat(i,"\n");flush.console()}

for(i in 1:(nrow(B)-1)){Lambdachf[i,(i+1):nrow(B)]=Lambdachf[i,i]}
```

```
# ic avec effet dépendant du temps
cut.points=unique(B$time[B$status2==1])
B2=survSplit(data=B,cut=cut.points,end="time",start="time0",event="status2")
B2$chft=B2$chf*B2$time
mCoxchf2=coxph(Surv(time0,time,status2)~chf+chft,data=B2)
summary(mCoxchf2)

der=B2$id[which.max(B2$time)]
intervalles=B2[B2$id==der,c("time0","time","status2")]

covs=data.frame(chf=1,intervalles)
covs$chft=covs$chf*covs$time

cumhaz=matrix(0,nrow=2,ncol=nrow(B))
cumhaz[1,1:length(survfit(mCoxchf2,newdata=data.frame(chf=0,chft=0))$cumhaz)]=
survfit(mCoxchf2,newdata=data.frame(chf=0,chft=0))$cumhaz
cumhaz[2,1:length(survfit(mCoxchf2,newdata=covs,individual=TRUE)$cumhaz)]=
survfit(mCoxchf2,newdata=covs,individual=TRUE)$cumhaz
for (i in 2:nrow(B)){if(cumhaz[1,i]==0){cumhaz[1,i]=cumhaz[1,i-1]}}
for (i in 2:nrow(B)){if(cumhaz[2,i]==0){cumhaz[2,i]=cumhaz[2,i-1]}}

Lambdachfdt=matrix(0,nrow=nrow(B),ncol=nrow(B))
for (i in 1:nrow(B)){
  if(B$chf[i]==0)
    Lambdachfdt[i,1:i]=cumhaz[1,1:i]
  else{Lambdachfdt[i,1:i]=cumhaz[2,1:i]}
  cat(i,"\n");flush.console()}

for(i in 1:(nrow(B)-1)){Lambdachfdt[i,(i+1):nrow(B)]=Lambdachfdt[i,i]}

# Figure 4
x11();par(mfrow=c(1,2))
plot(colSums(Lambdachf[which(B$chf==1),])[B$chf==1&B$status2==1]~cumsum(
summary(survfit(Surv(time,status2)~1,data=B,subset=(chf==1)))$n.event),
type="l",xlim=c(0,700),ylim=c(0,700),lty=2,lwd=5,col="grey0",xlab="Nombre
d'événements observés",ylab="Nombre d'événements estimés",cex.lab=1.7,
cex.axis=1.7)
lines(colSums(Lambdachf[which(B$chf==0),])[B$chf==0&B$status2==1]~cumsum(
summary(survfit(Surv(time,status2)~1,data=B,subset=(chf==0)))$n.event),
type="l",lty=2,lwd=5,col="grey70")
abline(0,1)
legend(0,700,c("Présence d'ic","Absence d'ic"),lty=2,lwd=5,col=c("grey0",
```

```
"grey70"),cex=1.7)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)

plot(colSums(Lambdachfdt[which(B$chf==1),])[B$chf==1&B$status2==1]~cumsum(
summary(survfit(Surv(time,status2)~1,data=B,subset=(chf==1)))$n.event),
type="l",xlim=c(0,700),ylim=c(0,700),lty=2,lwd=5,col="grey0",xlab="Nombre
d'événements observés",ylab="Nombre d'événements estimés",cex.lab=1.7,
cex.axis=1.7)
lines(colSums(Lambdachfdt[which(B$chf==0),])[B$chf==0&B$status2==1]~cumsum(
summary(survfit(Surv(time,status2)~1,data=B,subset=(chf==0)))$n.event),
type="l",lty=2,lwd=5,col="grey70")
abline(0,1)
legend(0,700,c("Présence d'ic","Absence d'ic"),lty=2,lwd=5,col=c("grey0",
"grey70"),cex=1.7)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)


# modèle de Cox ajusté sur l'âge
mCoxage=coxph(Surv(time,status!=0)~age,data=B)

Lambdaage=matrix(0,nrow=nrow(B),ncol=nrow(B))
for (i in 1:nrow(B)){
  Lambdaage[i,1:i]=survfit(mCoxage,
  newdata=data.frame(age=B$age[i]))$cumhaz[1:i]
  cat(i,"\n");flush.console()}

for(i in 1:(nrow(B)-1)){Lambdaage[i,(i+1):nrow(B)]=Lambdaage[i,i]}

#  modèle de Cox ajusté sur exp(age/100)
mCoxagee=coxph(Surv(time,status!=0)~I(exp(age/100)),data=B)

Lambdaagee=matrix(0,nrow=nrow(B),ncol=nrow(B))
for (i in 1:nrow(B)){
  Lambdaagee[i,1:i]=survfit(mCoxagee,
  newdata=data.frame(age=B$age[i]))$cumhaz[1:i]
  cat(i,"\n");flush.console()}

for(i in 1:(nrow(B)-1)){Lambdaagee[i,(i+1):nrow(B)]=Lambdaagee[i,i]}

# Figure 5
x11();par(mfrow=c(1,2))
B$agecat=cut(B$age,breaks=c(20,59.61,68.23,75.39,100))
plot(colSums(Lambdaage[which(B$agecat=="(20,59.6]"),])[B$agecat==
```

```
"(20,59.6]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(20,59.6]")))$n.event),type="l",xlim=c(0,400),
ylim=c(0,400),lty=2,lwd=5,col="grey0",xlab="Number of observed events",
ylab="Number of estimated events",cex.lab=1.7,,cex.axis=1.7)
lines(colSums(Lambdaage[which(B$agecat=="(59.6,68.2]"),])[B$agecat
=="(59.6,68.2]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(59.6,68.2]")))$n.event),type="l",lty=2,lwd=5,
col="grey30")
lines(colSums(Lambdaage[which(B$agecat=="(68.2,75.4]"),])[B$agecat==
"(68.2,75.4]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(68.2,75.4]")))$n.event),type="l",lty=2,lwd=5,
col="grey60")
lines(colSums(Lambdaage[which(B$agecat=="(75.4,100]"),])[B$agecat==
"(75.4,100]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(75.4,100]")))$n.event),type="l",lty=2,lwd=5,
col="grey90")
abline(0,1)
legend(0,400,c(expression("âge"<="59,6 ans"),"âge=]59,6;68,2] ans","âge=
(68,2;75,4] ans","âge>75,4 ans"),lty=2,lwd=5,col=c("grey0","grey30","grey60",
"grey90"),cex=1.7)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)

plot(colSums(Lambdaagee[which(B$agecat=="(20,59.6]"),])[B$agecat=="(20,59.6]"
&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,data=B,subset=(agecat
=="(20,59.6]")))$n.event),type="l",xlim=c(0,400),ylim=c(0,400),lty=2,lwd=5,
col="grey0",xlab="Number of observed events",ylab="Number of estimated events",
cex.lab=1.7,cex.axis=1.7)
lines(colSums(Lambdaagee[which(B$agecat=="(59.6,68.2]"),])[B$agecat=="(59.6,
68.2]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,data=B,
subset=(agecat=="(59.6,68.2]")))$n.event),type="l",lty=2,lwd=5,col="grey30")
lines(colSums(Lambdaagee[which(B$agecat=="(68.2,75.4]"),])[B$agecat=="(68.2,
75.4]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,data=B,
subset=(agecat=="(68.2,75.4]")))$n.event),type="l",lty=2,lwd=5,col="grey60")
lines(colSums(Lambdaagee[which(B$agecat=="(75.4,100]"),])[B$agecat=="(75.4,
100]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,data=B,
subset=(agecat=="(75.4,100]")))$n.event),type="l",lty=2,lwd=5,col="grey90")
abline(0,1)
legend(0,400,c(expression("âge"<="59,6 ans"),"âge=]59,6;68,2] ans",
"âge=(68,2;75,4] ans","âge>75,4 ans"),lty=2,lwd=5,col=c("grey0","grey30",
"grey60","grey90"),cex=1.7)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)

#modèle de Cox multivarié
```

```
cut.points=unique(B$time[B$status2==1])
B2=survSplit(data=B,cut=cut.points,end="time",start="time0",event="status2")
B2$chft=B2$chf*B2$time
B2$vft=B2$vf*B2$time
B2$vft2=B2$vf*(B2$time-0.15)*(B2$time>0.15)
mCoxm=coxph(Surv(time0,time,status2)~I(exp(age/100))+sex+chf+chft+dia+vf+
vft+vft2,data=B2)
summary(mCoxm)

testCoxm=cox.zph(mCoxm,transform="km")
testCoxm
x11();par(mfrow=c(2,4));plot(testCoxm)
x11();plot(testCoxm[5])

B2$diat=B2$dia*B2$time
mCoxm2=coxph(Surv(time0,time,status2)~I(exp(age/100))+sex+chf+chft+dia+diat+
vf+vft+vft2,data=B2)
summary(mCoxm2)
testCoxm2=cox.zph(mCoxm2,transform="km")
testCoxm2

# calcul de la somme des carrés des pseudo-résidus
StxCdt<-matrix(0,ncol=nrow(B),nrow=nrow(B))
if(is.list(mCoxm2)==T){
  der=B2$id[which.max(B2$time)]
  intervalles=B2[B2$id==der,c("time0","time","status2")]
  for (i in 1:nrow(B)){
    B3=data.frame(age=B$age[i],sex=B$sex[i],chf=B$chf[i],dia=B$dia[i],
    vf=B$vf[i],intervalles)
    B3$chft=B3$chf*B3$time
    B3$diat=B3$dia*B3$time
    B3$vft=B3$vf*B3$time
    B3$vft2=B3$vf*(B3$time-0.15)*(B3$time>0.15)
    StxCdt[i,]=survfit(mCoxm2,newdata=B3,individual=TRUE)$surv}
}

# Figure 6
# rapport des risques instantanés pour l'âge
x11()
j=function(x){exp(mCoxm2$coefficient[1]*exp(x/100))/exp(mCoxm2$coefficient[1]
*exp((x-1)/100))}
jinf=function(x){exp((mCoxm2$coefficient[1]-1.96*sqrt(mCoxm2$var[1,1]))*
exp(x/100))/exp((mCoxm2$coefficient[1]-1.96*sqrt(mCoxm2$var[1,1]))*
```

```
exp((x-1)/100))}
jsup=function(x){exp((mCoxm2$coefficient[1]+1.96*sqrt(mCoxm2$var[1,1]))*
exp(x/100))/exp((mCoxm2$coefficient[1]+1.96*sqrt(mCoxm2$var[1,1]))*
exp((x-1)/100))}
curve(j,24,97,ylim=c(0.95,1.1),cex.axis=1.7,lwd=3,cex.lab=1.7,xlab="âge
(ans)",ylab="Rapport des risques instantanés",cex.main=2,main="")
curve(jinf,add=T,lty=2,lwd=3)
curve(jsup,add=T,lty=2,lwd=3)
abline(h=1,lty=3)

# Figure 7
x11();
par(mfrow=c(1,3))
# rapport des risques instantanés pour l'ic
f=function(x){exp(mCoxm2$coefficient[3]+mCoxm2$coefficient[4]*x)}
finf=function(x){exp(mCoxm2$coefficient[3]+mCoxm2$coefficient[4]*x-1.96*
sqrt(mCoxm2$var[3,3]+x^2*mCoxm2$var[4,4]+2*x*mCoxm2$var[3,4]))}
fsup=function(x){exp(mCoxm2$coefficient[3]+mCoxm2$coefficient[4]*x+1.96*
sqrt(mCoxm2$var[3,3]+x^2*mCoxm2$var[4,4]+2*x*mCoxm2$var[3,4]))}
curve(f,0,8.5,ylim=c(0,4),cex.axis=1.7,lwd=3,cex.lab=2,cex.axis=2,xlab=
"temps (années)",ylab="",cex.main=3,main="chf")
curve(finf,add=T,lty=2,lwd=3)
curve(fsup,add=T,lty=2,lwd=3)
abline(h=1,lty=3)
mtext("Rapport des risques instantanés",side=2,adj=0.5,line=2.3,cex=1.7)

# rapport des risques instantanés pour le diabète
g=function(x){exp(mCoxm2$coefficient[5]+mCoxm2$coefficient[6]*x)}
ginf=function(x){exp(mCoxm2$coefficient[5]+mCoxm2$coefficient[6]*x-1.96*
sqrt(mCoxm2$var[5,5]+x^2*mCoxm2$var[6,6]+2*x*mCoxm2$var[5,6]))}
gsup=function(x){exp(mCoxm2$coefficient[5]+mCoxm2$coefficient[6]*x+1.96*
sqrt(mCoxm2$var[5,5]+x^2*mCoxm2$var[6,6]+2*x*mCoxm2$var[5,6]))}
curve(g,0,8.5,ylim=c(0,6),cex.axis=1.7,lwd=3,cex.lab=2,cex.axis=2,xlab=
"temps (années)",ylab="",cex.main=3,main="dia")
curve(ginf,add=T,lty=2,lwd=3)
curve(gsup,add=T,lty=2,lwd=3)
abline(h=1,lty=3)
mtext("Rapport des risques instantanés",side=2,adj=0.5,line=2.3,cex=1.7)

# rapport des risques instantanés pour la fv
h=function(x){exp(mCoxm2$coefficient[7]+x*mCoxm2$coefficient[8]+(x-0.15)*
mCoxm2$coefficient[9]*(x>0.15))}
hinf=function(x){exp(mCoxm2$coefficient[7]+x*mCoxm2$coefficient[8]+(x-0.15)*
```

```
mCoxm2$coefficient[9]*(x>0.15)-1.96*sqrt(mCoxm2$var[7,7]+x^2*mCoxm2$var[8,8]
+2*x*mCoxm2$var[7,8]+(x>0.15)*((x-0.15)^2*mCoxm2$var[9,9]+2*(x-0.15)*
mCoxm2$var[7,9]+2*x*(x-0.15)*mCoxm2$var[8,9]))))}
hsup=function(x){exp(mCoxm2$coefficient[7]+x*mCoxm2$coefficient[8]+(x-0.15)*
mCoxm2$coefficient[9]*(x>0.15)+1.96*sqrt(mCoxm2$var[7,7]+x^2*mCoxm2$var[8,8]
+2*x*mCoxm2$var[7,8]+(x>0.15)*((x-0.15)^2*mCoxm2$var[9,9]+2*(x-0.15)*
mCoxm2$var[7,9]+2*x*(x-0.15)*mCoxm2$var[8,9]))))}
curve(h,0,8.5,ylim=c(0,14),cex.axis=1.7,lwd=3,cex.lab=2,cex.axis=2,xlab="temps
(années)",ylab="",cex.main=3,main="vf")
curve(hinf,add=T,lty=2,lwd=3)
curve(hsup,add=T,lty=2,lwd=3)
abline(h=1,lty=3)
mtext("Rapport des risques instantanés",side=2,adj=0.5,line=2.3,cex=1.7)


# Proposition d'une stratégie de modélisation des modèles à risques instantanés
additifs
# Hypothèse de linéarité
# Figure 8
x11()
plot(-log(lowess(B$age,POS[,tps[1]],iter=0)$y)/times[tps[1]]~lowess(B$age,
POS[,tps[1]],iter=0)$x,xlab="age (years)",ylab="",type="l",ylim=c(-1,7),
cex.axis=1.7,lwd=3,cex.lab=1.7,
col="grey0")
points(68.2305,1.994219,pch=0,lwd=3,cex=1.5,col="grey0")
lines(-log(lowess(B$age,POS[,tps[2]],iter=0)$y)/times[tps[2]]~lowess(B$age,
POS[,tps[2]],iter=0)$x,type="l",col="grey10",lwd=3)
points(68.2305,1.061049,pch=1,lwd=3,cex=1.5,col="grey10")
lines(-log(lowess(B$age,POS[,tps[3]],iter=0)$y)/times[tps[3]]~lowess(B$age,
POS[,tps[3]],iter=0)$x,type="l",col="grey20",lwd=3)
points(68.2305,0.4040244,pch=2,lwd=3,cex=1.5,col="grey20")
lines(-log(lowess(B$age,POS[,tps[4]],iter=0)$y)/times[tps[4]]~lowess(B$age,
POS[,tps[4]],iter=0)$x,type="l",col="grey30",lwd=3)
points(68.2305,0.2114368,pch=3,lwd=3,cex=1.5,col="grey30")
lines(-log(lowess(B$age,POS[,tps[5]],iter=0)$y)/times[tps[5]]~lowess(B$age,
POS[,tps[5]],iter=0)$x,type="l",col="grey40",lwd=3)
points(68.2305,0.1600768,pch=4,lwd=3,cex=1.5,col="grey40")
lines(-log(lowess(B$age,POS[,tps[6]],iter=0)$y)/times[tps[6]]~lowess(B$age,
POS[,tps[6]],iter=0)$x,type="l",col="grey50",lwd=3)
points(68.2305,0.1425707,pch=5,lwd=3,cex=1.5,col="grey50")
lines(-log(lowess(B$age,POS[,tps[7]],iter=0)$y)/times[tps[7]]~lowess(B$age,
POS[,tps[7]],iter=0)$x,type="l",col="grey60",lwd=3)
points(68.2305,0.1209347,pch=6,lwd=3,cex=1.5,col="grey60")
```

```
lines(-log(lowess(B$age,POS[,tps[8]],iter=0)$y)/times[tps[8]]~lowess(B$age,
POS[,tps[8]],iter=0)$x,type="l",col="grey70",lwd=3)
points(68.2305,0.1097193,pch=7,lwd=3,cex=1.5,col="grey70")
lines(-log(lowess(B$age,POS[,tps[9]],iter=0)$y)/times[tps[9]]~lowess(B$age,
POS[,tps[9]],iter=0)$x,type="l",col="grey80",lwd=3)
points(68.2305,0.1069086,pch=8,lwd=3,cex=1.5,col="grey80")
abline(h=0)
legend("topleft",c("1er décile (0,03 an)","2e décile (0,09 an)",
"3e décile (0,37 an)","4e décile (0,97 an)","5e décile (1,73 an)",
"6e décile (2,46 ans)","7e décile (3,46 ans)","8e décile (4,61 ans)",
"9e décile (5,81 ans)"),col=c("grey0","grey9","grey18","grey27","grey36",
"grey45","grey54","grey63","grey72"),lwd=3,cex=1,pch=0:8,pt.cex=1.5)
mtext("-log(Pseudo-observations)/t",side=2,adj=0.5,line=2.5,cex=1.7)



# Processus de résidus de martingale
B$age2=cut(B$age,c(22.8,59.6,68.2,75.4,96.4),include.lowest=T)

MRP=function(event,covariates,contcat){
I=diag(as.numeric(event))
Z=rep(0,nrow(covariates)*(ncol(covariates)+1)*nrow(covariates))
dim(Z)=c(nrow(covariates),(ncol(covariates)+1),nrow(covariates))
Z[,1,]=1
for(i in 2:(ncol(covariates)+1)){
Z[,i,]=covariates[,i-1]}
for (i in 2:nrow(covariates)){Z[1:i-1,,i]=0}

beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(covariates))
for (i in 1:nrow(covariates)){
beta[,i]=tryCatch(solve(crossprod(Z[,,i],Z[,,i]))%*%t(Z[,,i])%*%I[i,],
warning=function(e) F,error=function(e) rep(0,ncol(Z)))}
Beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(covariates))
for(i in 1:dim(Z)[2]){Beta[i,]=cumsum(beta[i,])}

M=matrix(0,nrow=nrow(covariates),ncol=nrow(covariates))
J=diag(nrow(covariates))
for (i in 1:nrow(covariates)){
M[,i]=tryCatch((J-Z[,,i]%*%solve(crossprod(Z[,,i],Z[,,i]))%*%t(Z[,,i]))%*%
I[i,],warning=function(e) F,error=function(e) rep(0,nrow(covariates)))}
M2=matrix(0,nrow=nrow(covariates),ncol=nrow(covariates))
for(i in 1:nrow(covariates)){M2[i,]=cumsum(M[i,])}
```

```
model.matrix(~as.factor(contcat)-1)->contcatq
Mt<<-t(as.matrix(contcatq))%*%M2

# variance
V=rep(0,nrow(covariates)*nlevels(contcat)*nlevels(contcat))
dim(V)=c(nlevels(contcat),nlevels(contcat),nrow(covariates))
for (i in 1:nrow(covariates)){
V[,,i]=tryCatch(t(as.matrix(contcatq))%*%(J-Z[,,i]%*%solve(crossprod(Z[,,i],
Z[,,i]))%*%t(Z[,,i]))%*%diag(I[i,])%*%t(J-Z[,,i]%*%solve(crossprod(Z[,,i],
Z[,,i]))%*%t(Z[,,i]))%*%as.matrix(contcatq),warning=function(e) F,error=
function(e) matrix(0,nlevels(contcat),nlevels(contcat)))}
V2=rep(0,nrow(covariates)*nlevels(contcat)*nlevels(contcat))
dim(V2)=c(nlevels(contcat),nlevels(contcat),nrow(covariates))
for(i in 1:nlevels(contcat)){for (j in 1:nlevels(contcat)){V2[i,j,]<-
cumsum(V[i,j,])}}
V2<<-V2
for (i in 1:nlevels(contcat)){
print(1-pchisq(t(Mt[i,nrow(covariates)])%*%solve(V2[i,i,nrow(covariates)])
%*%Mt[i,nrow(covariates)],1))}
print(1-pchisq(t(Mt[1:(nlevels(contcat)-1),nrow(covariates)])%*%solve
(V2[1:(nlevels(contcat)-1),1:(nlevels(contcat)-1),nrow(covariates)])%*%
Mt[1:(nlevels(contcat)-1),nrow(covariates)],(nlevels(contcat)-1)))
}

MRP(event=B$status2,covariates=matrix(exp(B$age/10),nrow(B),ncol=1),
contcat=B$age2)
MRP(event=B$status2[1:100],covariates=matrix(B$age[1:100],100,ncol=1),
contcat=B$age2[1:100])

Mtb=Mt
V2b=V2
x11();par(mfrow=c(1,2))
plot(Mt[1,]~B$time,type="l",ylim=c(-80,90),xlab="temps (années)",
ylab="Processus de résidus de martingale",col="grey0",cex.axis=1.5,
cex.lab=1.5,lwd=3)
lines(Mt[2,]~B$time,type="l",lty=1,col="grey30",lwd=3)
lines(Mt[3,]~B$time,type="l",lty=1,col="grey60",lwd=3)
lines(Mt[4,]~B$time,type="l",lty=1,col="grey90",lwd=3)
legend(0,90,c(expression("âge"<="59,6 ans"),"âge=]59,6;68,2] ans",
"âge=(68,2;75,4] ans","âge>75,4 ans"),lty=1,lwd=5,col=c("grey0",
"grey30","grey60","grey90"),cex=1.7)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)
```

```
lines((Mt[1,]-1.96*sqrt(V2[1,1,]))~B$time,type="l",lty=2,col="grey0",lwd=3)
lines((Mt[1,]+1.96*sqrt(V2[1,1,]))~B$time,type="l",lty=2,col="grey0",lwd=3)
lines((Mt[2,]-1.96*sqrt(V2[2,2,]))~B$time,type="l",lty=2,col="grey30",lwd=3)
lines((Mt[2,]+1.96*sqrt(V2[2,2,]))~B$time,type="l",lty=2,col="grey30",lwd=3)
lines((Mt[3,]-1.96*sqrt(V2[3,3,]))~B$time,type="l",lty=2,col="grey60",lwd=3)
lines((Mt[3,]+1.96*sqrt(V2[3,3,]))~B$time,type="l",lty=2,col="grey60",lwd=3)
lines((Mt[4,]-1.96*sqrt(V2[4,4,]))~B$time,type="l",lty=2,col="grey90",lwd=3)
lines((Mt[4,]+1.96*sqrt(V2[4,4,]))~B$time,type="l",lty=2,col="grey90",lwd=3)
abline(h=0)

MRP(event=B$status2,covariates=matrix(c(exp(B$age/10),exp(B$age/10-7)*
(B$age>70)),nrow(B),ncol=2),contcat=B$age2)
plot(Mt[1,]~B$time,type="l",ylim=c(-80,90),xlab="temps (années)",ylab=
"Processus de résidus de martingale",col="grey0",cex.axis=1.5,cex.lab=1.5,
lwd=3)
lines(Mt[2,]~B$time,type="l",lty=1,col="grey30",lwd=3)
lines(Mt[3,]~B$time,type="l",lty=1,col="grey60",lwd=3)
lines(Mt[4,]~B$time,type="l",lty=1,col="grey90",lwd=3)
legend(0,90,c(expression("âge"<="59,6 ans"),"âge=]59,6;68,2] ans",
"âge=(68,2;75,4] ans","âge>75,4 ans"),lty=1,lwd=5,col=c("grey0",
"grey30","grey60","grey90"),cex=1.7)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)

lines((Mt[1,]-1.96*sqrt(V2[1,1,]))~B$time,type="l",lty=2,col="grey0",lwd=3)
lines((Mt[1,]+1.96*sqrt(V2[1,1,]))~B$time,type="l",lty=2,col="grey0",lwd=3)
lines((Mt[2,]-1.96*sqrt(V2[2,2,]))~B$time,type="l",lty=2,col="grey30",lwd=3)
lines((Mt[2,]+1.96*sqrt(V2[2,2,]))~B$time,type="l",lty=2,col="grey30",lwd=3)
lines((Mt[3,]-1.96*sqrt(V2[3,3,]))~B$time,type="l",lty=2,col="grey60",lwd=3)
lines((Mt[3,]+1.96*sqrt(V2[3,3,]))~B$time,type="l",lty=2,col="grey60",lwd=3)
lines((Mt[4,]-1.96*sqrt(V2[4,4,]))~B$time,type="l",lty=2,col="grey90",lwd=3)
lines((Mt[4,]+1.96*sqrt(V2[4,4,]))~B$time,type="l",lty=2,col="grey90",lwd=3)
abline(h=0)

# Figure 10
# covariable age
Amage=aareg(Surv(time,status!=0)~I(exp(age/10))+I(exp((age/10-7))*(age>70)),
data=B)
Linmage<-ahaz(Surv(B$time,B$status!=0),as.matrix(cbind(exp(B$age/10),
exp(B$age/10-7)*(B$age>70))))

# covariable sexe
Amsex=aareg(Surv(time,status!=0)~sex,data=B)
Linmsex<-ahaz(Surv(B$time,B$status!=0),as.matrix(B$sex))
```

```
# covariable ic
Amchf=aareg(Surv(time,status!=0)~chf,data=B)
Linmchf<-ahaz(Surv(B$time,B$status!=0),as.matrix(B$chf))

# covariable diabète
Amdia=aareg(Surv(time,status!=0)~dia,data=B)
Linmdia<-ahaz(Surv(B$time,B$status!=0),as.matrix(B$dia))

# covariate fv
Amvf=aareg(Surv(time,status!=0)~vf,data=B)
Linmvf<-ahaz(Surv(B$time,B$status!=0),as.matrix(B$vf))

x11();par(mfrow=c(2,3))

plot(Amage[2],cex.axis=2,cex.main=3,main="exp(age/10)",
col.lab="white")
abline(0,summary(Linmage)$coefficients[1,1],col=1,lty=2)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.2,cex=1.7)

plot(Amage[3],cex.axis=2,cex.main=3,main="exp(age/10-7)*(age>70)",
col.lab="white")
abline(0,summary(Linmage)$coefficients[2,1],col=1,lty=2)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(Amsex[2],cex.axis=2,cex.main=3,main="sexe",col.lab="white")
abline(0,summary(Linmsex)$coefficients[1,1],col=1,lty=2)
mtext("c",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(Amchf[2],cex.axis=2,cex.main=3,main="ic",col.lab="white",ylim=c(0,0.8))
abline(0,summary(Linmchf)$coefficients[1,1],col=1,lty=2)
mtext("d",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.2,cex=1.7)

plot(Amdia[2],cex.axis=2,cex.main=3,main="dia",col.lab="white",ylim=c(0,0.8))
abline(0,summary(Linmdia)$coefficients[1,1],col=1,lty=2)
```

```
mtext("e",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(Amvf[2],cex.axis=2,cex.main=3,main="fv",col.lab="white",ylim=c(0,0.8))
abline(0,summary(Linmvf)$coefficients[1,1],col=1,lty=2)
mtext("f",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)


# Figure 11
I=diag(as.numeric(B$status!=0))
Z=rep(0,nrow(B)*2*nrow(B))
dim(Z)=c(nrow(B),2,nrow(B))
Z[,1,]=1
Z[,2,]=B$age
for (i in 2:nrow(B)){
Z[1:i-1,,i]=0}

beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(B))
for (i in 1:nrow(B)){
beta[,i]=tryCatch(solve(crossprod(Z[,,i],Z[,,i]))%*%t(Z[,,i])%*%I[i,],
warning=function(e) F,error=function(e) rep(0,ncol(Z)))}
Beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(B))
for(i in 1:dim(Z)[2]){Beta[i,]=cumsum(beta[i,])}

Lambdaage=matrix(0,nrow=nrow(B),ncol=nrow(B))
for (i in 1:nrow(B)){for (j in 1:i){
Lambdaage[i,j]=Beta[1,j]+Beta[2,j]*B$age[i]}}
for(i in 1:(nrow(B)-1)){for(j in (i+1):nrow(B)){Lambdaage[i,j]=Lambdaage[i,i]}}

I=diag(as.numeric(B$status!=0))
Z=rep(0,nrow(B)*3*nrow(B))
dim(Z)=c(nrow(B),3,nrow(B))
Z[,1,]=1
Z[,2,]=exp(B$age/10)
Z[,3,]=exp((B$age-70)/10)*(B$age>70)
for (i in 2:nrow(B)){
Z[1:i-1,,i]=0}

beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(B))
for (i in 1:nrow(B)){
```

```
beta[,i]=tryCatch(solve(crossprod(Z[,,i],Z[,,i]))%*%t(Z[,,i])%*%I[i,],
warning=function(e) F,error=function(e) rep(0,ncol(Z)))}
Beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(B))
for(i in 1:dim(Z)[2]){Beta[i,]=cumsum(beta[i,])}

Lambdaagee=matrix(0,nrow=nrow(B),ncol=nrow(B))
for (i in 1:nrow(B)){for (j in 1:i){
Lambdaagee[i,j]=Beta[1,j]+Beta[2,j]*exp(B$age[i]/10)+Beta[3,j]*
exp((B$age[i]-70)/10)*(B$age[i]>70)}}
for(i in 1:(nrow(B)-1)){for(j in (i+1):nrow(B)){Lambdaagee[i,j]=
Lambdaagee[i,i]}}


x11();par(mfrow=c(1,2))
B$agecat=cut(B$age,breaks=c(20,59.61,68.23,75.39,100))
plot(colSums(Lambdaage[which(B$agecat=="(20,59.6]"),])[B$agecat=="(20,59.6]"&
B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,data=B,subset=
(agecat=="(20,59.6]")))$n.event),type="l",xlim=c(0,400),ylim=c(0,400),lty=2,
lwd=5,col="grey0",xlab="Nombre d'événements observés",ylab="Nombre d'événements
estimés",cex.lab=1.7,,cex.axis=1.7)
lines(colSums(Lambdaage[which(B$agecat=="(59.6,68.2]"),])[B$agecat==
"(59.6,68.2]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(59.6,68.2]")))$n.event),type="l",lty=2,lwd=5,
col="grey30")
lines(colSums(Lambdaage[which(B$agecat=="(68.2,75.4]"),])[B$agecat==
"(68.2,75.4]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(68.2,75.4]")))$n.event),type="l",lty=2,lwd=5,
col="grey60")
lines(colSums(Lambdaage[which(B$agecat=="(75.4,100]"),])[B$agecat==
"(75.4,100]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(75.4,100]")))$n.event),type="l",lty=2,lwd=5,
col="grey90")
abline(0,1)
legend(0,400,c(expression("âge"<="59,6 ans"),"âge=]59,6;68,2] ans",
"âge=(68,2;75,4] ans","âge>75,4 ans"),lty=2,lwd=5,col=c("grey0","grey30",
"grey60","grey90"),cex=1.7)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)

plot(colSums(Lambdaagee[which(B$agecat=="(20,59.6]"),])[B$agecat==
"(20,59.6]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(20,59.6]")))$n.event),type="l",xlim=c(0,400),
ylim=c(0,400),lty=2,lwd=5,col="grey0",xlab="Nombre d'événements observés",
ylab="Nombre d'événements estimés",cex.lab=1.7,cex.axis=1.7)
```

```
lines(colSums(Lambdaagee[which(B$agecat=="(59.6,68.2]"),])[B$agecat==
"(59.6,68.2]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(59.6,68.2]")))$n.event),type="l",lty=2,lwd=5,
col="grey30")
lines(colSums(Lambdaagee[which(B$agecat=="(68.2,75.4]"),])[B$agecat==
"(68.2,75.4]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(68.2,75.4]")))$n.event),type="l",lty=2,lwd=5,
col="grey60")
lines(colSums(Lambdaagee[which(B$agecat=="(75.4,100]"),])[B$agecat==
"(75.4,100]"&B$status2==1]~cumsum(summary(survfit(Surv(time,status2)~1,
data=B,subset=(agecat=="(75.4,100]")))$n.event),type="l",lty=2,lwd=5,
col="grey90")
abline(0,1)
legend(0,400,c(expression("âge"<="59,6 ans"),"âge=]59,6;68,2] ans",
"âge=(68,2;75,4] ans","âge>75,4 ans"),lty=2,lwd=5,col=c("grey0",
"grey30","grey60","grey90"),cex=1.7)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)

# modèle multivarié
multAm=aareg(Surv(time,status!=0)~I(exp(age/10))+I(exp((age/10-7))*
(age>70))+sex+chf+dia+vf,data=B)
summary(multAm)

multLinm=ahaz(Surv(B$time,B$status!=0),as.matrix(cbind(exp(B$age/10),
exp((B$age/10-7))*(B$age>70),B$sex,B$chf,B$dia,B$vf)))
summary(multLinm)

# Tableau 4
MRP(event=B$status2,covariates=as.matrix(cbind(exp(B$age/10),
exp((B$age/10-7))*(B$age>70),B$sex,B$chf,B$dia,B$vf)),contcat=B$age2)


# Figure 12
x11();par(mfrow=c(2,3))

plot(multAm[2],cex.axis=2,cex.main=3,main="exp(age/10)",col.lab="white")
abline(0,summary(multLinm)$coefficients[1,1],col=1,lty=2)
mtext("a",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.2,cex=1.7)

plot(multAm[3],cex.axis=2,cex.main=3,main="exp(age/10-7)*(age>70)",
col.lab="white")
```

```
abline(0,summary(multLinm)$coefficients[2,1],col=1,lty=2)
mtext("b",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(multAm[4],cex.axis=2,cex.main=3,main="sexe",col.lab="white",ylim=c(0,0.8))
abline(0,summary(multLinm)$coefficients[3,1],col=1,lty=2)
mtext("c",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(multAm[5],cex.axis=2,cex.main=3,main="ic",col.lab="white",ylim=c(0,0.8))
abline(0,summary(multLinm)$coefficients[4,1],col=1,lty=2)
mtext("d",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.2,cex=1.7)

plot(multAm[6],cex.axis=2,cex.main=3,main="dia",col.lab="white",ylim=c(0,0.8))
abline(0,summary(multLinm)$coefficients[5,1],col=1,lty=2)
mtext("e",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

plot(multAm[7],cex.axis=2,cex.main=3,main="fv",col.lab="white",ylim=c(0,0.8))
abline(0,summary(multLinm)$coefficients[6,1],col=1,lty=2)
mtext("f",side=1,adj=0,line=3.5,cex=2.5)
mtext("temps (années)",side=1,adj=0.5,line=3.5,cex=1.7)
mtext("Alpha(t)",side=2,adj=0.5,line=2.5,cex=1.7)

# calcul de la somme des carrés des pseudo-résidus
# matrice I
I=diag(B$status!=0)

# matrice Z
Z=rep(0,nrow(B)*7*nrow(B))
dim(Z)=c(nrow(B),7,nrow(B))
Z[,1,]=1
Z[,2,]=exp(B$age/10)
Z[,3,]=exp((B$age/10-7))*(B$age>70)
Z[,4,]=B$sex
Z[,5,]=B$chf
Z[,6,]=B$dia
Z[,7,]=B$vf
```

```
for (i in 2:nrow(B)){
  Z[1:i-1,,i]=0}

beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(B))
for (i in 1:nrow(B)){
  beta[,i]=tryCatch(solve(crossprod(Z[,,i],Z[,,i]))%*%t(Z[,,i])%*%I[i,],
  warning=function(e) F,error=function(e) rep(0,ncol(Z)))}
Beta=matrix(0,nrow=dim(Z)[2],ncol=nrow(B))
for(i in 1:dim(Z)[2]){Beta[i,]=cumsum(beta[i,])}
StxAq=exp(-Z[,,1]%*%Beta)

resCSbrut=POS-StxCdt
resASbrut=POS-StxAq
c(sum(resCSbrut^2),sum(resASbrut^2))


# Figure 13
x11()
f=function(x){(summary(multLinm)$coefficients[1,1]*exp(x/10)+
summary(multLinm)$coefficients[2,1]*exp(x/10-7)*(x>70))}
curve(f,2.3,9.6)

x11()
j=function(x){(summary(multLinm)$coefficients[1,1]*exp(x/10)+
summary(multLinm)$coefficients[2,1]*exp(x/10-7)*(x>70))}
jinf=function(x){(summary(multLinm)$coefficients[1,1]-1.96*
summary(multLinm)$coefficients[1,2])*exp(x/10)+(summary(multLinm)$
coefficients[2,1]-1.96*summary(multLinm)$coefficients[2,2])*
exp(x/10-7)*(x>70)}
jsup=function(x){(summary(multLinm)$coefficients[1,1]+1.96*
summary(multLinm)$coefficients[1,2])*exp(x/10)+(summary(multLinm)$
coefficients[2,1]+1.96*summary(multLinm)$coefficients[2,2])*
exp(x/10-7)*(x>70)}
curve(j,23,96,ylim=c(0,2),cex.axis=1.7,lwd=3,cex.lab=1.7,xlab=
"âge (années)",ylab="Alpha (t)",cex.main=2,main="")
curve(jinf,add=T,lty=2,lwd=3)
curve(jsup,add=T,lty=2,lwd=3)
abline(h=1,lty=3)
```

C. Figures annexes

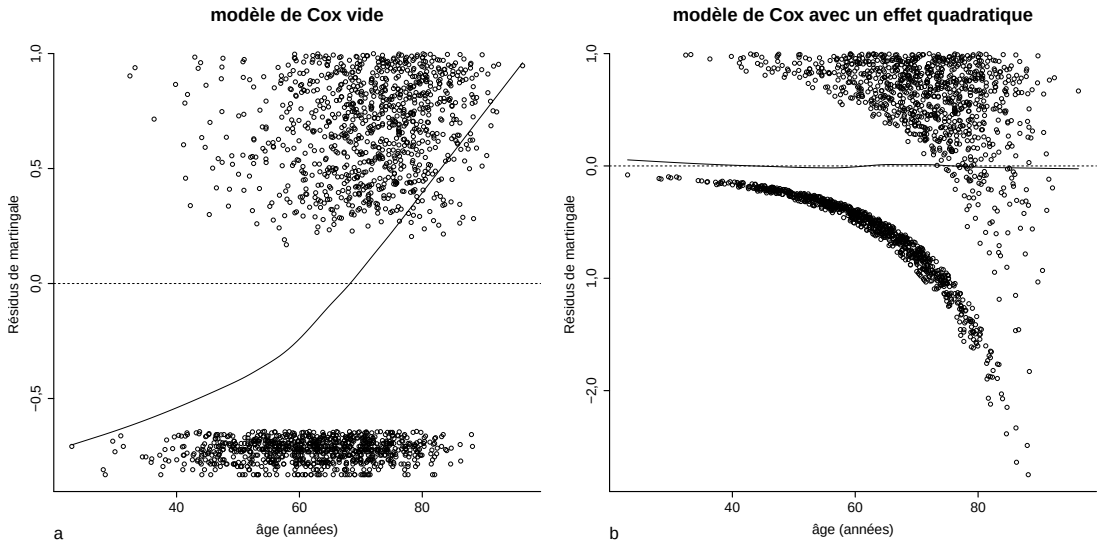


FIGURE .1. – Résidus de martingale en fonction de l'âge modélisé avec (a) un modèle de Cox vide et avec (b) un modèle de Cox ajusté avec une fonction quadratique de l'âge sur 100. Une courbe de lissage est ajoutée en trait plein sur chaque graphique.

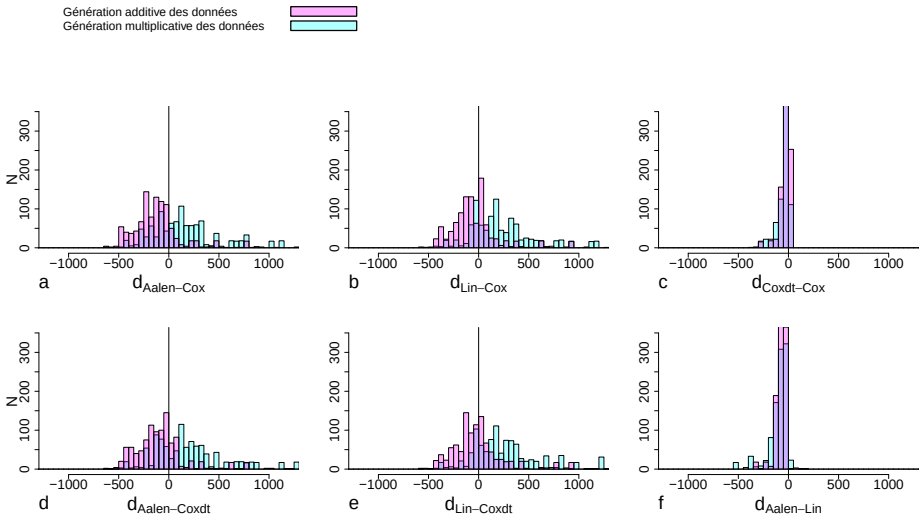


FIGURE .2. – Distributions des différences des SCPR estimées obtenues avec les quatre modèles à risques instantanés multiplicatifs ou additifs comparés deux à deux. Situation avec une variable générée continue constante et d'effet moyen, une censure de 30 % et un échantillon de 1 000 sujets.

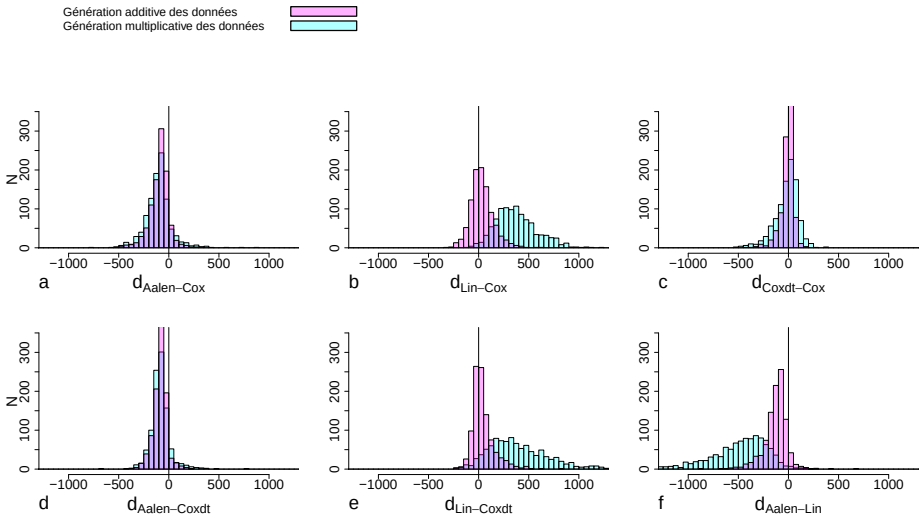


FIGURE .3. – Distributions des différences des SCPR estimées obtenues avec les quatre modèles à risques instantanés multiplicatifs ou additifs comparés deux à deux. Situation avec une variable générée binaire constante et d'effet moyen, une censure de 30 % et un échantillon de 1000 sujets.

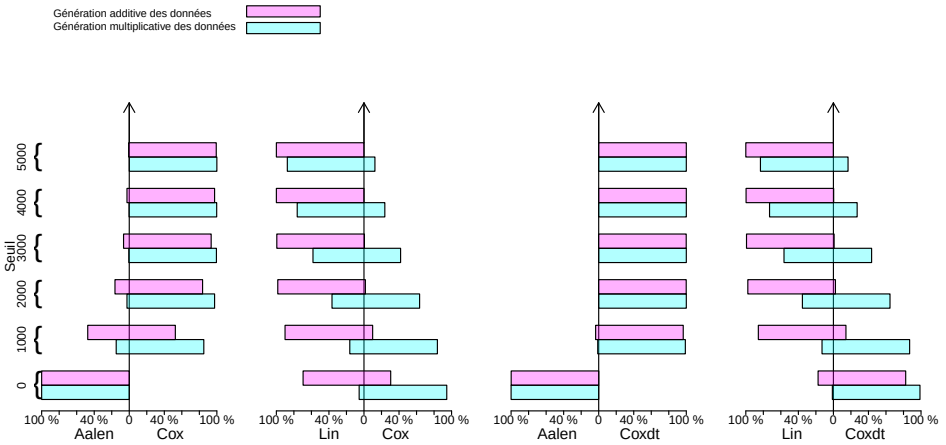


FIGURE .4. – Pourcentage des cas pour lesquels le meilleur modèle s'ajustant le mieux aux données entre un modèle à risques instantanés additifs (Aalen, Lin) et multiplicatifs (Cox, Coxdt) était sélectionné en fonction de la valeur du seuil et du modèle de génération des données. Situation avec une variable générée binaire constante et d'effet moyen, et un échantillon de 1 000 sujets.

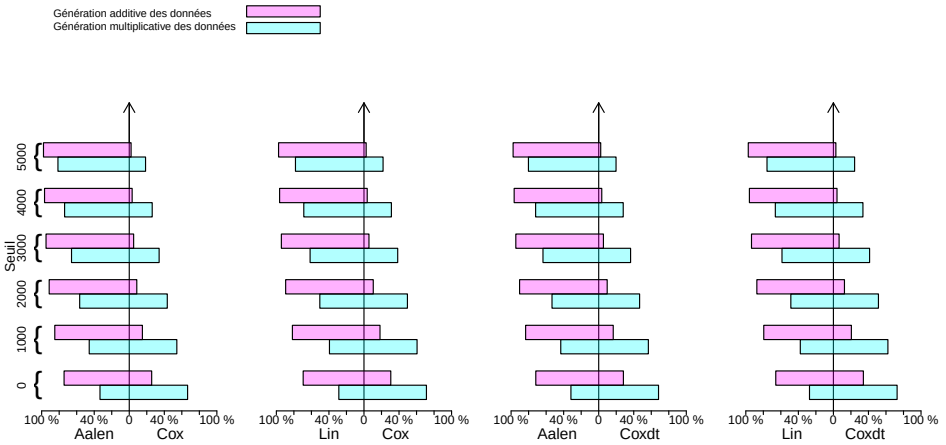


FIGURE .5. – Pourcentage des cas pour lesquels le meilleur modèle s'ajustant le mieux aux données entre un modèle à risques instantanés additifs (Aalen, Lin) et multiplicatifs (Cox, Coxdt) était sélectionné en fonction de la valeur du seuil et du modèle de génération des données. Situation avec une variable générée continue constante et d'effet moyen, et un échantillon de 1 000 sujets.