

EVE-MARIE CASTONGUAY

**MODÉLISATION DE LA SURVIE RELATIVE :
APPLICATION AUX ACCIDENTS
VASCULAIRES CÉRÉBRAUX**

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

Faculté des sciences et de génie
UNIVERSITÉ LAVAL
QUÉBEC

Décembre 2004

©Eve-Marie Castonguay, 2004

Résumé

La survie relative est définie comme le rapport entre la survie observée dans un groupe de sujets atteints d'une certaine maladie et la survie attendue pour un groupe comparable de la population générale, formé de sujets non atteints de la maladie. L'objectif de ce mémoire est de comparer différentes méthodes d'estimation de la survie relative. Un survol des principales méthodes de calcul de la survie observée et de la survie attendue est d'abord proposé. Vient ensuite une comparaison mathématique de plusieurs approches de modélisation de la survie relative, toutes basées sur la maximisation de la vraisemblance. Enfin, une comparaison pratique des approches de modélisation est suggérée. Celle-ci est basée sur les données d'une étude de type populationnel sur les accidents vasculaires cérébraux.

Avant-Propos

Je voudrais d'abord et avant tout remercier mon directeur de recherche, monsieur Louis-Paul Rivest, professeur au Département de mathématiques et de statistique de l'Université Laval, qui, en plus de ses précieux conseils, a toujours su se montrer disponible et m'a encouragée tout au long de ce travail. Je remercie aussi monsieur Belkacem Abdous, professeur au Département de médecine sociale et préventive de l'Université Laval, qui a proposé et co-supervisé ce projet.

J'aimerais également exprimer toute ma gratitude à tous les membres de l'Unité connaissance-surveillance de l'Institut National de Santé Publique du Québec, avec qui j'ai passé le plus clair de mon temps pendant ma maîtrise. Plus particulièrement, merci à madame Danielle St-Laurent, coordonatrice, qui m'a chaleureusement accueillie dans son équipe, m'a encouragée, supportée, conseillée et qui m'a volontiers fourni un espace de travail agréable et pratique. J'aimerais aussi adresser un merci particulier à madame Rabiâ Louchini, épidémiologiste, dont les connaissances sur la survie relative ainsi que sur les AVC m'ont été d'une aide précieuse. Rabiâ a toujours pris le temps de répondre à mes questions et de me prodiguer de précieux conseils.

Merci aussi à ma mère, à Isabelle, à Rabiâ et à Josianne. Elles ont accepté de relire mon mémoire et m'ont suggéré quelques idées et corrections.

Mais je voudrais par dessus tout remercier ma famille, sans qui la poursuite de mes études n'aurait pu être possible, mes amis, qui m'ont fait rire et m'ont permis de me divertir même dans les moments les plus difficiles et évidemment mon amoureux des cinq dernières années, Jean-François, qui m'a épaulée et encouragée à persévérer tout au long de mes études universitaires. Je dédie donc ce mémoire à tous ces gens que j'aime et que j'adore et qui, peut-être sans le savoir, ont contribué de près ou de loin à la réussite de mon projet.

Ce travail de recherche a été financé en partie par l'Institut National de Santé Publique du Québec.

Table des matières

Résumé	ii
Avant-Propos	iii
Liste des tableaux	vii
Table des figures	viii
Introduction	1
1 Notions préliminaires	3
1.1 Fonctions liées à la survie	4
1.1.1 La fonction de survie	4
1.1.2 La fonction de risque	5
1.2 Rappel sur la loi de Poisson et les modèles linéaires généralisés	7
1.2.1 Loi de Poisson	7
1.2.2 Modèles linéaires généralisés	8
1.3 Relation entre le taux de mortalité et la probabilité de survie	8
2 Survie relative	10
2.1 Méthodes classiques de calcul de la survie observée	14
2.1.1 Fonction de survie empirique	14
2.1.2 La méthode actuarielle	14
2.1.3 La méthode de Kaplan-Meier	15
2.1.4 Comparaison des méthodes actuarielle et de Kaplan-Meier	16
2.2 Méthodes d'estimation de la survie attendue	17
2.2.1 Méthode Ederer I	17
2.2.2 Méthode Ederer II	18
2.2.3 Méthode d'Hakulinen	19
2.2.4 Comparaison des méthodes Ederer I, Ederer II et Hakulinen	23
3 Modélisation de la survie relative	25

3.1	Méthode d'Estève <i>et al.</i> (1990) : approche par maximisation de la vraisemblance	25
3.2	Étude d'un cas particulier du modèle d'Estève <i>et al.</i>	27
3.3	Extension du modèle d'Estève <i>et al.</i> :	
	modélisation Poisson	30
	3.3.1 Extension du modèle d'Estève <i>et al.</i> :	
	modélisation binomiale	33
4	Application aux accidents vasculaires cérébraux	35
4.1	Mise en situation	35
4.2	Description de la base de données	36
4.3	Exploration des données	38
4.4	Estimation de la survie relative	40
	4.4.1 Survie observée	40
	4.4.2 Survie attendue	41
	4.4.3 Survie relative	42
4.5	Modélisation de la survie relative	44
4.6	Discussion	54
	Conclusion	57
	Bibliographie	59
	A Tables québécoises de mortalité de 1986 et 1991	64
	B Liste des variables provenant du fichier sur les accidents vasculaires cérébraux	68
	C Programme SAS pour la modélisation de la survie relative	69

Liste des tableaux

2.1	Estimation par la méthode Ederer I de la probabilité de survie attendue cumulative à 5 ans pour 15 sujets.	18
2.2	Estimation par la méthode Ederer II de la probabilité de survie attendue cumulative à 5 ans pour 15 sujets.	20
2.3	Estimation par la méthode d'Hakulinen de la proportion de survie attendue cumulative à 5 ans pour 15 sujets.	23
3.1	Table de mortalité pour l'individu i de la strate k , décédé en 1992 à 76 ans, 3.25 ans après le diagnostic ($t_i = 3.25, \delta_i = 1$).	28
3.2	Intervalles de suivi pour l'individu i de la strate k , décédé en 1992 à 76 ans, 3.25 ans après le diagnostic.	31
4.1	Nombre de cas d'AVC et nombre de décès de toutes causes parmi ces cas survenus dans la période de deux ans suivant l'hospitalisation, pour les années 1990 à 1992.	39
4.2	Nombre de cas d'AVC et nombre de décès de toutes causes parmi ces cas survenus dans la période de deux ans suivant l'hospitalisation, pour chacune des catégories d'âge.	39
4.3	Survie observée (%) spécifique aux intervalles (Int) et cumulative (Cum), par groupe d'âge et par sexe.	41
4.4	Estimation de la survie attendue spécifique aux intervalles (Int) et cumulative (Cum), par sexe, à l'aide des méthodes Ederer I et Ederer II.	41
4.5	Survie relative (%) spécifique aux intervalles (Int) et cumulative (Cum), par groupe d'âge et par sexe.	44
4.6	Estimations des ratios d'excès de risque ($\exp(\beta)$) pour quatre approches de modélisation de la survie relative pour les cas d'AVC survenus entre 1990 et 1992.	49
4.7	Étapes pour l'ajustement du modèle de survie relative choisi.	50
4.8	Statistiques du rapport de vraisemblances pour l'analyse de Type 3.	51
4.9	Estimations des paramètres et des ratios d'excès de risque ($\exp(\beta)$) pour le modèle final.	52

Table des figures

1.1	Distribution de la survie après le diagnostic d'angine de poitrine. Le temps 0 correspond au diagnostic d'angine de poitrine.	6
4.1	Survie attendue cumulative calculée à l'aide de la méthode Ederer II, par catégories d'âge pour chacun des intervalles de suivi.	43
4.2	Vérification graphique de l'hypothèse de proportionnalité des risques pour la variable âge lors de l'hospitalisation.	46

Introduction

L'analyse des durées de vie est un domaine de la statistique qui étudie l'apparition d'un événement au cours du temps. Pour ce faire, il est nécessaire de disposer du temps de suivi de tous les individus à l'étude, ainsi que du moment auquel l'événement s'est produit (dans le cas où évidemment celui-ci a eu lieu). Ce qui est particulier avec ce type d'étude, c'est la présence de données censurées (données incomplètes) pour les sujets chez qui l'événement d'intérêt est non observé. L'analyse de ces données nécessite ainsi une méthodologie adaptée. L'analyse des durées de vie est donc particulièrement utile pour étudier plusieurs types d'événements, notamment des bris d'équipement, des tremblements de terre, des divorces et évidemment, des décès.

Les premières analyses statistiques d'une durée de vie remonteraient au *XVII^e* siècle. La notion de survie relative, représentant le rapport entre la survie observée et la survie attendue, apparut beaucoup plus tard, en 1942, et fut introduite par [Berkson \(1942\)](#). En 1950, [Berkson et Gage \(1950\)](#) définirent de façon plus précise le concept, sans toutefois suggérer une méthode pratique de calcul. Leurs travaux ont eu comme résultat de fournir une mesure objective de la proportion de patients qui décèdent des conséquences directes ou indirectes d'une maladie dans une population donnée. Par conséquent, ces travaux ont fourni une mesure corrigée de la survie de patients pour les effets des autres causes indépendantes de décès. Les premières méthodes de calcul univariées et reposant sur des statistiques non paramétriques furent introduites par [Ederer et Heise \(1959\)](#), [Ederer *et al.* \(1961\)](#) et [Hakulinen \(1982\)](#). Ces méthodes se nomment respectivement Ederer II, Ederer I et Hakulinen et reposent sur le calcul de la survie attendue. Ensuite, des approches fondées sur la modélisation du taux ont fait leur apparition. [Breslow \(1975\)](#) a été le premier à proposer une modélisation multiplicative, qui considère un taux relatif par rapport à un taux de base. Ce modèle a vite été critiqué ([Buckley, 1984](#)) et d'autres modèles multiplicatifs ont été développés, notamment par [Andersen et Vaeth \(1989\)](#). Par contre, plusieurs auteurs ([Andersen et Vaeth, 1989](#); [Buckley, 1984](#); [Hakulinen et Tenkanen, 1987](#)) ont jugé plus naturel d'employer une modélisation additive. En effet, de tels modèles sont généralement plus plausibles du

point de vue biologique et donnent un meilleur ajustement des données, surtout pour les données de survie de type populationnel. Vint ensuite la modélisation multivariée, qui a permis de combler les besoins d'estimation de la survie, en fonction de plusieurs facteurs pronostiques, et ce, à partir de modèles statistiquement puissants. Encore une fois, des modèles multiplicatifs (Breslow *et al.*, 1983; Andersen *et al.*, 1985) et additifs (Pocock *et al.*, 1982; Zahl, 1993, 1995; Hakulinen et Tenkanen, 1987; Estève *et al.*, 1990; Marubini *et al.*, 1990; Chevart et Ryan, 1991; Sasiemi, 1996) ont été suggérés. Toutefois, les modèles additifs sont plus utilisés, étant plus faciles d'interprétation.

C'est une étude de type populationnel sur les accidents vasculaires cérébraux, amorcée à l'Institut National de Santé Publique du Québec (INSPQ), qui a motivé ce mémoire et qui a, par le fait même, suscité un intérêt pour la survie relative. Cette méthode est fréquemment employée pour tenir compte des risques différentiels de décès dans les études sur le cancer de type populationnel.

Le chapitre I fait d'abord un tour d'horizon des définitions et concepts de base nécessaires à la compréhension du mémoire. Le chapitre II présente la notion de survie relative, ainsi que différentes méthodes de calcul de la survie observée et de la survie attendue. Sont ensuite exposés, dans le chapitre III, les principaux modèles additifs de survie relative, qui sont tous des extensions du modèle de survie relative d'Estève *et al.* (1990). Ces différentes méthodes d'estimation et de modélisation seront finalement illustrées et comparées au chapitre IV, à l'aide des données sur les accidents vasculaires cérébraux fournies par l'Institut National de Santé Publique du Québec.

Chapitre 1

Notions préliminaires

L'analyse des durées de vie réfère aux méthodes employées pour l'étude du temps jusqu'à l'occurrence d'un certain événement, comme un décès ou l'apparition de symptômes. Nous appelons temps de survie les données qui mesurent le délai entre le début de l'étude et la manifestation de l'événement.

Les méthodes standards d'analyse statistique sont inappropriées pour les données de survie. En effet, ces données présentent plusieurs particularités, notamment une distribution non symétrique. Pour cette raison, de nouveaux modèles de distribution doivent être adoptés. De plus, les temps de survie sont fréquemment censurés. Par définition, le temps de survie d'un individu est dit censuré, lorsque sa valeur exacte n'est pas observée ; seules des bornes supérieures ou inférieures pour cette valeur sont disponibles. La censure peut se manifester pour différentes raisons : l'événement d'intérêt n'est pas survenu au moment de l'analyse, un sujet peut être perdu de vue avant d'avoir expérimenté l'événement d'intérêt, un événement concurrent peut être survenu avant l'événement d'intérêt, un sujet peut être exclu de l'étude sans avoir expérimenté l'événement d'intérêt, etc.

Il existe plusieurs types de censure dont la censure à droite, la censure à gauche et la censure par intervalle. Seule la censure à droite peut être considérée pour l'application des méthodes présentées dans ce mémoire.

La censure à droite est la forme de censure la plus commune dans les études médicales. Ce type de censure survient lorsque l'événement d'intérêt se produit après

la fin de la période de suivi. En d'autres mots, il y a une borne inférieure pour la durée de vie. Cela peut, par exemple, se produire lors d'un essai clinique pour tester la survie de souris à un nouveau virus. Certaines souris décèdent entre le moment de l'injection du virus et la date de la fin de l'étude, tandis que d'autres survivront jusqu'à la fin. La durée de vie des souris non décédées, à la fin de l'étude, sera censurée à droite.

1.1 Fonctions liées à la survie

La distribution des temps de survie est généralement caractérisée par trois fonctions : la fonction de survie, la fonction de densité et la fonction de risque. Ces trois fonctions sont, en fait, interreliées, car la connaissance d'une seule est suffisante pour dériver les autres. Les sections suivantes présentent ces trois fonctions liées à la survie.

1.1.1 La fonction de survie

Nous noterons T ($T \geq 0$), la variable aléatoire qui représente la durée de vie d'une unité dans une expérience.

La fonction de survie, notée $S(t)$, est définie comme la probabilité qu'un individu survive au-delà du temps t , c'est-à-dire qu'il expérimente l'événement après le temps t :

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

Lorsque T est une variable aléatoire continue, la fonction de survie est le complément de la fonction de répartition, car

$$S(t) = 1 - P(\text{un individu échoue avant le temps } t) = 1 - F(t),$$

où $F(t) = P(T \leq t)$.

De plus, si la dérivée de $S(t)$ existe, la variable aléatoire T a une densité $f(t)$ estimée comme le rapport entre les individus décédés dans l'intervalle et la largeur de l'intervalle. Ainsi, la fonction de survie est l'intégrale de la fonction de densité $f(t)$:

$$S(t) = P(T > t) = \int_t^{\infty} f(x) dx.$$

Donc,

$$f(t) = \frac{-dS(t)}{dt}.$$

La fonction de survie $S(t)$ possède trois caractéristiques importantes :

- $S(t)$ est une fonction monotone croissante ;
- $S(t) = 1$ pour $t = 0$;
- $S(t) = 0$ pour $t = \infty$.

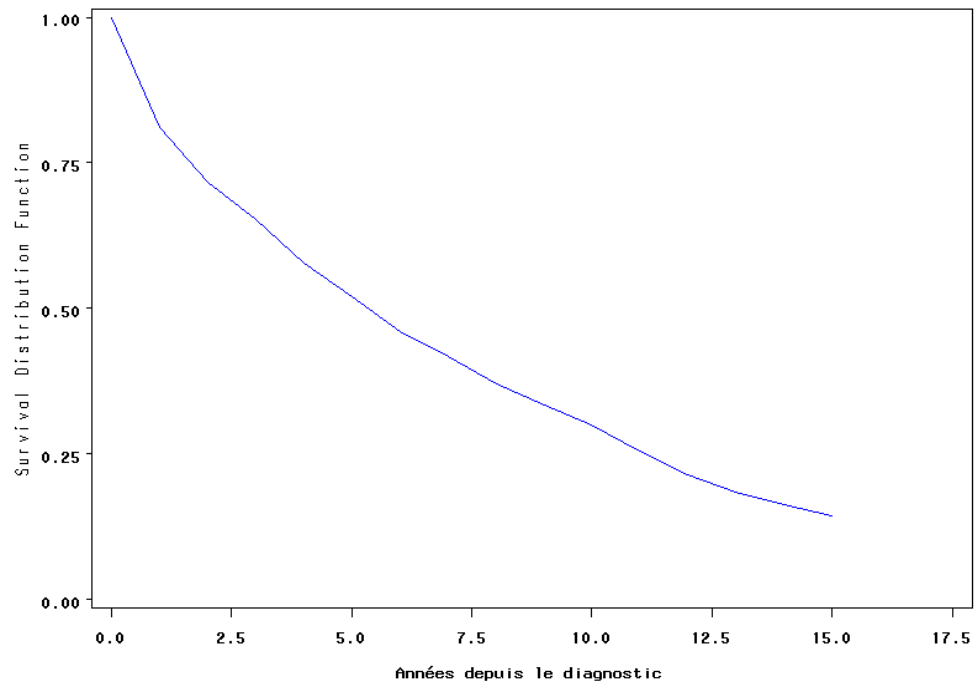
En somme, la probabilité de survivre au moins au temps 0 est de 1 et la probabilité de survivre un temps infini est de 0.

La représentation graphique de la survie, appelée courbe de survie, est une fonction monotone non croissante de la probabilité de survie en fonction du temps. Son taux de déclin varie selon le risque d'expérimenter l'événement au temps t . La figure 1.1 est un exemple de courbe de survie. Elle représente la survie d'hommes souffrant d'angine de poitrine (voir [Aide en ligne de SAS \(1999\)](#), procédure LIFETEST, exemple 37.2). Le temps de survie est mesuré en années depuis le diagnostic.

1.1.2 La fonction de risque

La fonction de risque, aussi appelée taux de panne, taux de décès conditionnel ou

FIG. 1.1 – Distribution de la survie après le diagnostic d'angine de poitrine. Le temps 0 correspond au diagnostic d'angine de poitrine.



force de mortalité, est la proportion de décès dans un intervalle par unité de temps, parmi les sujets encore vivants au début de l'intervalle. Elle est définie par

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t}.$$

Si T est une variable aléatoire continue,

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-d \ln[S(t)]}{dt}.$$

La fonction de risque est une mesure de propension au décès en fonction de l'âge de l'individu. En d'autres termes, cette fonction donne un risque de décès par unité de temps tout au long du processus de vieillissement. Les épidémiologistes ont baptisé cette fonction taux de mortalité.

La fonction de risque donne habituellement plus d'informations sur le mécanisme de décès que la fonction de survie. C'est pourquoi cette fonction est souvent employée pour résumer les données de survie.

La fonction de risque cumulé est définie par :

$$\Lambda(t) = \int_0^t \lambda(x) dx = -\ln[S(t)]$$

et son domaine se situe entre 0 et l'infini.

Par ailleurs, différentes méthodes peuvent être employées pour estimer la survie. Quelques-unes de ces méthodes seront présentées au chapitre suivant.

1.2 Rappel sur la loi de Poisson et les modèles linéaires généralisés

1.2.1 Loi de Poisson

Une variable aléatoire Y est dite de distribution Poisson de paramètre λ si elle prend des valeurs positives $y = 0, 1, 2, \dots$ avec probabilité

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad \text{si } \lambda > 0.$$

Il peut être montré que la moyenne et la variance de cette distribution sont

$$E(Y) = \text{Var}(Y) = \lambda.$$

La log-vraisemblance associée au paramètre λ d'une loi de Poisson, basée sur un échantillon de variables aléatoires Y_1, \dots, Y_n indépendantes et identiquement distribuées, est donnée par

$$l(\lambda) = -n\lambda + \sum_{i=1}^n y_i \log(\lambda).$$

1.2.2 Modèles linéaires généralisés

Le modèle linéaire généralisé est une extension du modèle linéaire, car il s'adapte à la fois aux réponses de distribution non normales et aux transformations pour la linéarité. Un modèle linéaire généralisé se définit de la façon suivante :

- Il existe un ensemble de variables réponses Y_1, \dots, Y_n , qui suivent une distribution provenant de la famille exponentielle, appelé composante aléatoire. La composante aléatoire peut donc être de nature discrète ou continue.
- La composante systématique spécifie une combinaison linéaire de variables explicatives utilisées comme prédicteurs dans le modèle, c'est-à-dire $\alpha + \beta_1 x_1 + \dots + \beta_k x_k$, où les x_i ont une influence sur la distribution de la réponse.
- La moyenne μ est une fonction du prédicteur linéaire. L'inverse de cette fonction est appelée fonction de lien. La fonction de lien $g(\mu)$ décrit comment $\mu = E(Y)$ lie les variables explicatives au prédicteur linéaire. Cette fonction de lien est $g(\mu) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$. Les fonctions de lien les plus connues sont les suivantes : logarithmique, logit et probit.
- Une variable offset est une variable explicative ajoutée à la composante systématique du modèle dont le coefficient est 1.

L'objectif du modèle linéaire généralisé est en fait de modéliser la fonction monotone $g(\mu)$. Cela peut se faire en estimant les paramètres α et β par la méthode du maximum de vraisemblance.

1.3 Relation entre le taux de mortalité et la probabilité de survie

La mortalité est habituellement caractérisée par un taux. Ce taux, tout simplement appelé taux de mortalité, est le rapport entre le nombre de décès survenus au cours d'une période dans une population donnée à l'effectif de cette population au milieu de la période ([Bouchard et Louchini, 1999](#)). En revanche, une table de mortalité présente

une probabilité de décéder à chaque âge dans une population donnée, en fonction du taux de mortalité. Il existe donc une relation entre le taux de mortalité, la probabilité de décès et la probabilité de survie.

Supposons que l'individu i est atteint d'une certaine maladie et que son âge, lors du diagnostic de cette maladie, était de a_i années. Supposons de plus que l'étude de la survie de cet individu repose sur des temps de suivi prédéterminés $t_0 < t_1 < \dots < t_J$, où $t_0 = 0$ et J est le nombre d'intervalles de suivi, et que les t_j sont exprimés en années. S'il est encore vivant au début du j ème intervalle de suivi, cet individu sera âgé de $a_{ji} = a_i + t_{j-1}$ années. Si, dans la population, le taux de mortalité pour un individu d'âge a_{ji} années est de $\lambda(a_{ji})$, la probabilité qu'il survive jusqu'à la fin du j ème intervalle de suivi, en sachant qu'il était vivant au début de cet intervalle, est de

$$p_j(i) = [1 - \lambda(a_{ji})]^{(t_j - t_{j-1})}. \quad (1.1)$$

En effet, pour un intervalle de temps j de longueur $(t_j - t_{j-1})$, la probabilité de décès $1 - p_j(i)$, pour l'individu i à l'âge a_{ji} , peut être modélisée par une loi exponentielle de paramètre $\lambda(a_{ji})$:

$$1 - p_j(i) = 1 - \exp[-(t_j - t_{j-1})\lambda(a_{ji})].$$

Cette relation peut aussi s'écrire

$$p_j(i)^{1/(t_j - t_{j-1})} = \exp[-\lambda(a_{ji})].$$

À l'aide d'un développement en série de Taylor, nous obtenons approximativement

$$p_j(i) = [1 - \lambda(a_{ji})]^{(t_j - t_{j-1})}.$$

Nous pouvons donc considérer, pour les besoins de ce mémoire, que $p_j(i) = [1 - \lambda(a_{ji})]^{(t_j - t_{j-1})}$ pour des longueurs d'intervalles inférieures à 1 et que $p_j(i) = [1 - \lambda(a_{ji})]$ pour des intervalles de longueur 1.

Chapitre 2

Survie relative

Lorsque nous étudions la survie d'un groupe de patients atteints d'une certaine maladie, nous devons distinguer deux concepts : la survie brute et la survie nette.

La survie brute mesure la survie pour l'ensemble des forces de mortalité, y compris la force de mortalité liée à la maladie étudiée. Par contre, la survie nette représente la survie pour la maladie à l'étude dans la situation hypothétique où toutes les autres causes de décès seraient éliminées. Pour ce faire, nous devons supposer que la cause spécifique de décès étudiée agit indépendamment de toutes les autres causes de décès. La survie nette peut être estimée par la méthode de la survie spécifique, qui ne prend en compte que les décès associés à la maladie étudiée. Cette méthode est en fait rarement utilisable, car elle nécessite de connaître la cause exacte du décès. Cela est souvent rendu difficile par l'imprécision des certificats de décès et la faible fréquence des autopsies ([Foucher et al., 1994](#)). La méthode de la survie spécifique peut aussi être interprétée en termes de risques concurrents, mais cette façon de faire comporte des obstacles méthodologiques importants ([Crowder, 2001](#)). C'est pourquoi une deuxième méthode d'estimation de la survie nette a été proposée. Cette méthode, appelée survie relative ([Berkson et Gage, 1950](#)), n'impose pas de connaître la cause exacte du décès comme c'est le cas avec la survie spécifique. La survie relative se définit comme étant le rapport entre la survie observée dans un groupe de sujets atteints d'une certaine maladie et la survie prédite pour un groupe comparable de la population générale formé de sujets non atteints de la maladie.

Le concept de survie relative est apparu dans les années 50 afin de tenir compte des risques différentiels de décès, ce que ne permet pas le modèle semi-paramétrique de [Cox](#)

(1972). En fait, l'idée de base du modèle de Cox, voulant qu'un taux de base arbitraire a priori inconnu doit être estimé, a été reprise pour un modèle de risques additifs (Klein et Moeschberger, 1997, chap 10), plutôt que multiplicatifs. Dans ce cas-ci, le taux de base est estimé à l'aide de tables de mortalité. Le modèle additif utilisé s'écrit en termes de $\lambda(t; \mathbf{x}; a)$, la force de mortalité au temps t d'un sujet d'âge a années au diagnostic, où \mathbf{x} est un vecteur de covariables pouvant éventuellement influencer la survie. L'équation de base est

$$\lambda(t; \mathbf{x}; a) = \lambda^*(a + [t]; \mathbf{x}) + \nu(t; \mathbf{x}), \quad (2.1)$$

où $\nu(t; \mathbf{x})$ est la force de mortalité spécifique à la maladie d'intérêt, aussi appelée excès de risque dû à la présence de la maladie, et $\lambda^*(a + [t]; \mathbf{x})$ est la force de mortalité associée aux autres causes de décès. $\lambda^*(a + [t]; \mathbf{x})$ est une fonction en escalier constante sur des intervalles de type $[j, j + 1]$, pour $j \in \mathbb{N}$.

Plus précisément, $\lambda^*(a + [t]; \mathbf{x})$ est la fonction de risque connue, ou attendue, à partir des tables de mortalité. Cette force de mortalité est donc estimée à partir de données externes, qui sont les taux de mortalité de la population générale. Par convention, un astérisque annote le risque attendu pour préciser qu'il est estimé à partir des taux de mortalité de l'ensemble de la population. Ce risque dépend donc de l'année du diagnostic, ainsi que de l'âge pour chacun des individus. Par exemple, pour un homme de 75 ans qui a reçu un diagnostic de cancer en 1987 et qui est décédé dans la même année, $\lambda^*(a + [t]; \mathbf{x})$ sera égal à la probabilité qu'un homme de 75 ans décède selon les tables québécoises de mortalité, qui sont publiées tous les 5 ans.

Par ailleurs, ce modèle peut s'écrire sous la forme suivante :

$$S(t; \mathbf{x}; a) = S^*(a + [t]; \mathbf{x})r(t; \mathbf{x}), \quad (2.2)$$

où $S(t; \mathbf{x}; a)$, $S^*(a + [t]; \mathbf{x})$ et $r(t; \mathbf{x}; a)$ représentent respectivement la survie observée cumulative, la survie attendue et la survie relative.

En effet, si nous intégrons de part et d'autre de l'équation (2.1), nous obtenons

$$\int_0^t \lambda(s; \mathbf{x}; a) ds = \int_0^t [\lambda^*(a + [s]; \mathbf{x}) + \nu(s; \mathbf{x})] ds.$$

Ainsi,

$$\exp\left(-\int_0^t \lambda(s; \mathbf{x}; a) ds\right) = \exp\left(-\int_0^t \lambda^*(a + [s]; \mathbf{x}) ds\right) \exp\left(-\int_0^t \nu(s; \mathbf{x}) ds\right),$$

ce qui se traduit par

$$S(t; \mathbf{x}; a) = S^*(a + [t]; \mathbf{x}) r(t; \mathbf{x})$$

et donc

$$r(t; \mathbf{x}) = \frac{S(t; \mathbf{x}; a)}{S^*(a + [t]; \mathbf{x})} = \frac{\text{Survie observée}}{\text{Survie attendue}} = \frac{p}{p^*},$$

où $r(t; \mathbf{x})$ est appelé survie relative.

De plus, si nous supposons que la composante d'excès de risque $\nu(t; \mathbf{x})$ est une fonction multiplicative de covariables qui ne dépend pas de t , $\exp(\mathbf{x}\beta)$, le modèle de survie relative de base s'écrit :

$$\lambda(t; \mathbf{x}; a) = \lambda^*(a + [t]; \mathbf{x}) + \exp(\mathbf{x}\beta). \quad (2.3)$$

Au temps t , où t est le temps depuis le diagnostic, la survie relative peut être interprétée comme la proportion de sujets malades qui auraient survécu au moins au temps t et ce, dans la situation hypothétique où la maladie en question serait la seule

cause possible de décès.

Pour calculer la survie relative, les individus du groupe de comparaison servant au calcul de la survie attendue doivent présenter les mêmes caractéristiques individuelles (âge, sexe, race) que les individus à l'étude, sans toutefois être atteints de la maladie d'intérêt. En fait, cette méthode repose sur l'hypothèse que le groupe de patients atteints de la maladie est soumis à deux forces de mortalité : la mortalité liée à la maladie à l'étude et la mortalité liée à toutes les autres causes de décès.

En outre, l'information sur la cause réelle de décès n'est pas nécessaire. Seulement la date du décès est requise. Par conséquent, l'emploi de la survie relative permet d'éviter les problèmes d'imprécision ou de non-disponibilité des certificats de décès ainsi que de l'incertitude quant à la cause du décès. Le calcul de la survie relative demande seulement de savoir que le sujet a eu un diagnostic de maladie et qu'il était vivant ou mort à un moment précis.

De plus, la stabilisation de la survie relative, qui se traduit par l'horizontalité de la courbe de survie relative, est souvent interprétée en termes de guérison. Si, par exemple, nous retrouvons dans une population un rapport de survie relative stable, après un certain nombre d'années de suivi, cela indique qu'une partie du groupe de patients aurait échappé à la force de mortalité de la maladie à l'étude. [Hédelin \(2000\)](#) suggère par contre d'être critique face à cette interprétation, notamment en ce qui concerne une surveillance importante de la maladie étudiée. En effet, la surveillance pourrait alors jouer le rôle de dépistage, impliquant donc la découverte précoce des pathologies. Ainsi, la mortalité par autres causes des individus atteints de la maladie à l'étude pourrait devenir plus faible que celle de la population générale. La période au bout de laquelle les individus auraient échappé à la force de mortalité de la maladie à l'étude pourrait donc être sous-estimée.

Pour estimer le ratio de survie relative, il faut d'abord calculer la survie observée dans l'échantillon et ensuite estimer la survie attendue. Les sections suivantes présentent chacune des méthodes pouvant être employées pour le calcul de la survie observée et l'estimation de la survie attendue.

2.1 Méthodes classiques de calcul de la survie observée

2.1.1 Fonction de survie empirique

Supposons d'abord un échantillon simple de temps de survie, où nous ne retrouvons pas d'observations censurées. La fonction de survie $S(t)$ est la probabilité qu'un individu survive pour un temps supérieur ou égal à t . Cette fonction peut être estimée par la fonction de survie empirique, donnée par

$$\tilde{S}(t) = \frac{\text{Nombre d'individus avec des temps de survie } \geq t}{\text{Nombre d'individus à l'étude}}.$$

De façon équivalente, $\tilde{S}(t) = 1 - \tilde{F}(t)$, où $\tilde{F}(t)$ est la fonction de répartition empirique, c'est-à-dire le rapport entre le nombre total d'individus en vie au temps t et le nombre total d'individus dans l'étude. Il est à noter que $\tilde{S}(t)$ est égale à 1 pour les valeurs de t précédant le premier temps de décès et est égale à 0 après le dernier temps de décès.

Cette méthode d'estimation de la fonction de survie ne peut malheureusement pas être employée lorsque nous sommes en présence de temps de survie censurés. Il existe par contre deux méthodes non-paramétriques très connues pour l'estimation de $S(t)$ en présence de censure, c'est-à-dire la méthode actuarielle et la méthode de Kaplan-Meier (aussi appelée méthode du produit-limite).

2.1.2 La méthode actuarielle

Cette approche consiste à diviser la période d'observation en une série d'intervalles de temps et à estimer la proportion de survie conditionnelle pour chaque intervalle. Les intervalles de temps sont fixés a priori. En absence de censure, la proportion de survie pour un intervalle spécifique est

$$p = \frac{(l - d)}{l},$$

où d est le nombre d'événements observés pendant l'intervalle et l est le nombre de sujets en vie au début de celui-ci. En présence de censure à droite, il est convenu d'assumer que celle-ci survient de façon aléatoire dans l'intervalle. Chaque donnée censurée dans un intervalle contribue alors d'un demi-intervalle au temps à risque de décès cumulé sur cet intervalle. Ce principe est appelé hypothèse actuarielle. Le nombre de sujets à risque sur tout l'intervalle est donc donné par

$$l' = l - \frac{w}{2},$$

où w est le nombre de temps de survie censurés pendant cette période.

L'estimation de la proportion de survie observée, pour un intervalle spécifique, est donc

$$p = \frac{(l' - d)}{l'}. \quad (2.4)$$

2.1.3 La méthode de Kaplan-Meier

Cet estimateur standard de la fonction de survie a été proposé par [Kaplan et Meier \(1958\)](#). Au temps t , cet estimateur est défini comme suit :

$$\hat{S}(t) = \begin{cases} 1 & \text{si } t < t_1 \\ \prod_{t_i \leq t} (1 - \frac{d_i}{l_i}) & \text{si } t \geq t_1 \end{cases}$$

où l_i est le nombre d'unités à risque avant les décès du temps t_i et d_i est le nombre de décès qui surviennent au temps t_i .

La méthode de Kaplan-Meier repose sur les mêmes principes que la méthode actuarielle. Par contre, les intervalles de temps sont déterminés a posteriori par les moments de décès observés : les probabilités conditionnelles de survie entre deux dates de décès sont estimées à chaque fois qu'un décès survient. Comme pour la méthode actuarielle, la probabilité de survie, à partir du début du suivi, est obtenue par le produit des probabilités calculées pour des intervalles successifs. De plus, une donnée censurée (à droite) dans l'intervalle est censurée au début de l'intervalle. La contribution au temps à risque de cette donnée sur l'intervalle est donc nulle. Par conséquent, la censure n'affecte pas l'estimation de $S(t)$, mais y contribue en faisant diminuer le nombre de sujets à risque au temps de décès suivant.

Une importante propriété de cet estimateur de la fonction de survie est qu'il est asymptotiquement distribué comme une loi normale de moyenne $S(t)$.

Par ailleurs, la variance de l'estimateur de Kaplan-Meier pour chaque temps t est estimée par la formule de Greenwood,

$$v[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{l_i(l_i - d_i)}.$$

2.1.4 Comparaison des méthodes actuarielle et de Kaplan-Meier

Dans les études de type populationnel, la méthode actuarielle est souvent préférée à la méthode de Kaplan-Meier et ce, même si les estimations de $S(t)$ produites par ces deux méthodes sont semblables. La principale raison de cela est que la méthode actuarielle utilise des données groupées, alors que la méthode de Kaplan-Meier ne s'appuie sur aucun regroupement de données. Puisque les données issues d'études de type populationnel sont souvent groupées, le calcul de la survie observée repose généralement sur la méthode actuarielle.

2.2 Méthodes d'estimation de la survie attendue

2.2.1 Méthode Ederer I

Ederer *et al.* (1961) ont été les premiers à proposer une méthode d'estimation de la survie attendue. Tout d'abord, le temps depuis le diagnostic est divisé en J intervalles de suivi. Une proportion de survie attendue est ensuite déterminée pour chacun de ces intervalles. Nous devons pour cela disposer d'une table de mortalité spécifique à chacune des strates k , où le nombre maximal de strates correspond au nombre de combinaisons possibles des différents facteurs qui affectent la survie, c'est-à-dire l'âge, le sexe, la race, etc.

Selon cette méthode, la proportion de survie attendue cumulative, à partir de la date de diagnostic jusqu'à la fin du J ème intervalle de temps, est donnée par :

$$p_J^* = \sum_{i=1}^n \frac{p_J^*(i)}{n}$$

et

$$p_J^*(i) = \prod_{j=1}^J p_j^*(i)$$

où

n est le nombre total de patients en vie au début du suivi ;

$p_j^*(i)$ est la probabilité de survie attendue à la fin de l'intervalle j pour l'individu i , à l'âge a_{ji} , et est définie à l'aide de la relation 1.1 ;

$p_J^*(i)$ est la probabilité de survie attendue à la fin du J ième intervalle pour un sujet de la population générale comparable au sujet i en vie au début du suivi.

En fait, nous calculons le produit des probabilités attendues des intervalles de suivi pour chaque individu. Ensuite, la moyenne de ces produits est établie pour chaque strate k .

Ainsi, la proportion de survie attendue à 5 ans est estimée par la moyenne des probabilités individuelles de survie attendue à 5 ans. Le tableau 2.1 présente un exemple de calcul de survie attendue à l'aide de cette méthode. Les probabilités de survie attendue sont tirées des tables de mortalité pour 1986 et 1991 présentées à l'annexe A. Puisque 1986 et 1991 sont les années centrales, les probabilités de survie attendue pour les années de diagnostic 1984, 1985, 1986, 1987 et 1988 sont tirées de la table de 1986, alors que les probabilités de survie attendue pour les années 1989, 1990, 1991, 1992 et 1993 sont tirées de la table de mortalité de 1991.

TAB. 2.1 – Estimation par la méthode Ederer I de la probabilité de survie attendue cumulative à 5 ans pour 15 sujets.

					Probabilités de survie attendue dans l'intervalle j					Survie attendue cumulative
ID	Sexe	Âge au diagnostic	Année du diagnostic	Temps de survie	1	2	3	4	5	$p_5^*(i)$
1	M	72	89	0	0.95700	0.95295	0.94865	0.94404	0.93904	0.76694
2	F	82	84	0	0.93206	0.92456	0.91646	0.90764	0.90413	0.64809
3	M	63	88	0	0.97797	0.97938	0.97757	0.97553	0.97316	0.88889
4	F	83	85	1	0.92456	0.91646	0.90764	0.90413	0.90501	0.62928
5	F	79	86	2	0.95030	0.94491	0.93888	0.94029	0.93364	0.74012
6	M	70	86	2	0.95795	0.95410	0.95004	0.95295	0.94865	0.78497
7	F	70	88	2	0.98051	0.98124	0.97915	0.97677	0.97416	0.89639
8	F	85	84	3	0.90764	0.90413	0.89370	0.88229	0.87002	0.56296
9	M	80	84	4	0.90592	0.89793	0.88939	0.88031	0.87067	0.55452
10	F	52	89	6	0.99678	0.99647	0.99615	0.99579	0.99538	0.98072
11	M	60	88	7	0.98341	0.98434	0.98271	0.98106	0.97938	0.91401
12	F	71	87	8	0.97853	0.97618	0.97677	0.97416	0.97126	0.88280
13	M	58	87	8	0.98622	0.98487	0.98588	0.98434	0.98271	0.92629
14	F	80	87	8	0.94491	0.93888	0.94029	0.93364	0.92645	0.72155
15	F	56	86	9	0.99464	0.99414	0.99359	0.99372	0.99305	0.96952
Estimation de la probabilité de survie attendue cumulative sur 5 ans (p_5^*)										0.79114

2.2.2 Méthode Ederer II

Ederer et Heise (1959) ont proposé une méthode alternative à la méthode Ederer I pour estimer la survie attendue. Cette méthode permet, contrairement à la première, de considérer des temps de suivi hétérogènes. La proportion de survie attendue cumulative se définit comme le produit de la survie attendue moyenne calculé pour chaque intervalle de temps de suivi et s'écrit :

$$p_J^* = \prod_{j=1}^J p_{j2}^*,$$

où

$$p_{j2}^* = \sum_{i=1}^{n_j} \frac{p_j^*(i)}{n_j}$$

est la moyenne des probabilités de survie attendue ($p_j^*(i)$), définies par l'équation 1.1, des sujets en vie au début de l'intervalle j (n_j).

En fait, les proportions de survie attendue pour les intervalles spécifiques sont estimées pour les J intervalles et ce, à partir des patients vivants au début de l'intervalle seulement (n_j). La proportion de survie attendue cumulative est donc estimée par le produit des proportions de survie moyennes d'intervalles spécifiques. Un exemple de cette méthode pour les mêmes 15 individus de la table 2.1 est présenté au tableau 2.2, où les probabilités de survie attendue sont tirées des tables de mortalité présentées à l'annexe A.

2.2.3 Méthode d'Hakulinen

Dans la méthode d'Hakulinen (1982), comme dans la méthode Ederer II, nous calculons la survie attendue au début des J intervalles de temps pour les personnes encore en vie. Dans cette méthode, nous ajoutons en plus une composante permettant de tenir compte du temps potentiel de suivi des individus. Pour des sujets ayant un temps de suivi censuré, le temps potentiel de suivi est le temps de suivi observé, alors que pour

TAB. 2.2 – Estimation par la méthode Ederer II de la probabilité de survie attendue cumulative à 5 ans pour 15 sujets.

ID	Sexe	Âge au diagnostic	Année du diagnostic	Temps de survie	Probabilités de survie attendue dans l'intervalle j				
					1	2	3	4	5
1	M	72	89	0	0.95700				
2	F	82	84	0	0.93206				
3	M	63	88	0	0.97797				
4	F	83	85	1	0.92456	0.91646			
5	F	79	86	2	0.95030	0.94491	0.93888		
6	M	70	86	2	0.95795	0.95410	0.95004		
7	F	70	88	2	0.98051	0.98124	0.97915		
8	F	85	84	3	0.90764	0.90413	0.89370	0.88229	
9	M	80	84	4	0.90592	0.89793	0.88939	0.88031	0.87067
10	F	52	89	6	0.99678	0.99647	0.99615	0.99579	0.99538
11	M	60	88	7	0.98341	0.98434	0.98271	0.98106	0.97938
12	F	71	87	8	0.97853	0.97618	0.97677	0.97416	0.97126
13	M	58	87	8	0.98622	0.98487	0.98588	0.98434	0.98271
14	F	80	87	8	0.94491	0.93888	0.94029	0.93364	0.92645
15	F	56	86	9	0.99464	0.99414	0.99359	0.99372	0.99305
Probabilités de survie attendue des intervalles spécifiques (p_{j2}^*)					0.95856	0.95614	0.95696	0.95316	0.95984
Probabilité de survie attendue à 5 ans (p_5^*)					0.80242				

les sujets décédés, le temps potentiel de suivi est le temps pendant lequel ils auraient été suivis si le décès ne s'était pas produit (Therneau et Offord, 1999). Contrairement aux deux autres méthodes, celle-ci tient compte du nombre attendu de cas censurés pour chacun des intervalles de temps et de l'évolution de la structure de la cohorte observée. Dans les études de type populationnel, la censure se traduit habituellement par des pertes au suivi, par exemple l'émigration, ou par la fin de l'étude avant le moment de l'événement.

L'idée derrière cette méthode est d'introduire un estimateur biaisé de la proportion de survie attendue avec un biais similaire à celui présent dans les proportions de survie observée. Ainsi, le rapport de ces deux quantités résulte en un estimateur non biaisé du ratio de survie relative. Par contre, le principal inconvénient est que l'information sur les temps de suivi potentiels est requise pour tous les patients, ce qui est souvent difficile à obtenir (surtout pour les patients décédés).

Pour le calcul de la proportion de survie attendue par la méthode d'Hakulinen (Voutilainen *et al.*, 2000), nous noterons

- $K_j = K_{ja} + K_{jb}$ l'ensemble des patients avec un temps de suivi potentiel s'étendant au-delà du commencement du j ème intervalle ;
- K_{ja} l'ensemble des patients avec un temps de suivi potentiel s'étendant au-delà de la fin du j ème intervalle ;
- K_{jb} l'ensemble des censures potentielles pendant le j ème intervalle ;
- $p_j^*(i)$ la probabilité de survie attendue pour les i sujets en vie au début de l'intervalle j , telle que définie à la section 1.3 ;
- l_1 le nombre de patients au début du premier intervalle. l_1 joue le rôle du n des méthodes Ederer I et II.

Soit k_j le nombre de patients avec un temps de suivi potentiel s'étendant au-delà du commencement du j ème intervalle. Soit maintenant le premier k_{ja} de ces k_j patients avec un temps de suivi potentiel s'étendant au-delà de la fin du j ème intervalle et le dernier k_{jb} un retrait potentiel pendant le j ème intervalle. Il s'ensuit que $k_1 = l_1, k_{j+1} = k_{ja}$ et $k_j = k_{ja} + k_{jb}$. La notation K_{ja} sera employée pour désigner l'ensemble des k_{ja} patients, K_{jb} pour désigner l'ensemble des k_{jb} patients, etc.

Le nombre attendu de sujets en vie et sous observation au début du j ème intervalle est donné par :

$$l_j^* = \begin{cases} \sum_{i \in K_j} p_{j-1}^*(i) & \text{pour } j \geq 2 \\ l_1 & \text{pour } j = 1. \end{cases}$$

Pour les patients k_{jb} avec temps de suivi potentiels se terminant pendant le j ème intervalle, il est convenu que chacun des patients est à risque pour la moitié de l'intervalle, ce qui fait que la probabilité attendue de décès pendant cet intervalle est de $1 - \sqrt{p_j^*}$. En effet, sous l'hypothèse d'une distribution uniforme des censures et d'une distribution exponentielle des temps de survie pendant le j ème intervalle de temps, la probabilité d'être exclu vivant est proche de $(p_j^*)^{1/2}$ (Chiang, 1968). Il est donc possible de calculer le nombre d'individus pour lesquels le temps de suivi est compris dans le j ème intervalle. Le nombre attendu de patients qui se retirent de l'étude, mais qui sont en vie pendant le j ème intervalle, est donc donné par :

$$w_j^* = \begin{cases} \sum_{i \in K_{jb}} p_{j-1}^*(i) \sqrt{p_j^*(i)} & \text{pour } j \geq 2 \\ \sum_{i \in K_{1b}} \sqrt{p_1^*(i)} & \text{pour } j = 1. \end{cases}$$

Le nombre attendu de patients décédant pendant le j ème intervalle, parmi les k_{jb} patients avec temps de suivi potentiels se terminant pendant le même intervalle, est donné par :

$$\delta_j^* = \begin{cases} \sum_{i \in K_{jb}} p_{j-1}^*(i) [1 - \sqrt{p_j^*(i)}] & \text{pour } j \geq 2 \\ \sum_{i \in K_{1b}} [1 - \sqrt{p_1^*(i)}] & \text{pour } j = 1 \end{cases}$$

et le nombre attendu total de sujets à décéder pendant le j ème intervalle est donné par :

$$d_j^* = \begin{cases} \{\sum_{i \in K_{ja}} p_{j-1}^*(i) [1 - p_j^*(i)]\} + \delta_j^* & \text{pour } j \geq 2 \\ \{\sum_{i \in K_{1a}} [1 - p_1^*(i)]\} + \delta_1^* & \text{pour } j = 1. \end{cases}$$

La proportion de survie attendue pour un intervalle spécifique peut donc s'écrire :

$$g_j^* = 1 - \frac{d_j^*}{l_j^* - w_j^*/2}$$

et, finalement, la proportion de survie attendue depuis le début du suivi jusqu'à la fin du J ème intervalle est obtenue en calculant :

$$p_J^* = \prod_{j=1}^J g_j^*.$$

Le tableau 2.3 reprend l'exemple présenté pour les méthodes Ederer I et II, et présente les différentes composantes qui ont été nécessaires au calcul de la survie attendue par la méthode d'Hakulinen. Par exemple, si on illustre les calculs pour le troisième intervalle, on a

- $K_3 = 11$, le nombre total de sujets en vie au début du troisième intervalle ;
- $K_{3a} = 11$, le nombre total de sujets en vie au début du quatrième intervalle ;

- $K_{3b} = 0$, le nombre d'individus avec un temps de suivi potentiellement censuré durant le troisième intervalle (il n'y a pas de données censurées dans cet exemple) ;
- $l_3^* = \sum_{i \in K_3} p_2^*(i) = 0.94491 + 0.95410 + \dots + 0.99414 = 10.55719$;
- $w_3^* = \sum_{i \in K_{3b}} p_2^*(i) \sqrt{p_3^*(i)} = 0$;
- $\delta_3^* = \sum_{i \in K_{3b}} p_2^*(i) [1 - \sqrt{p_3^*(i)}] = 0$;
- $d_3^* = \{ \sum_{i \in K_{3a}} p_2^*(i) [1 - p_3^*(i)] \} + \delta_3^* = \{ 0.94491[1 - 0.93888] + 0.95410[1 - 0.95004] + \dots + 0.99414[1 - 0.99359] \} + 0 = 0.44118$;
- $g_3^* = 1 - \frac{d_3^*}{l_3^* - w_3^*/2} = 1 - \frac{0.44118}{10.55719 - 0} = 0.95821$,

où toutes les probabilités de survie attendue sont tirées du tableau 2.2.

TAB. 2.3 – Estimation par la méthode d'Hakulinen de la proportion de survie attendue cumulative à 5 ans pour 15 sujets.

Intervalle	K_j	K_{ja}	K_{jb}	l_j^*	w_j^*	δ_j^*	d_j^*	g_j^*
1	15	15	0	15	0	0	0.62160	0.95856
2	12	12	0	11.51137	0	0	0.49217	0.95724
3	11	11	0	10.55719	0	0	0.44118	0.95821
4	8	8	0	7.65848	0	0	0.34377	0.95511
5	7	7	0	6.74302	0	0	0.25927	0.96155
Estimation de la proportion de survie attendue à 5 ans (p_5^*)								0.80747

2.2.4 Comparaison des méthodes Ederer I, Ederer II et Hakulinen

La méthode Ederer I produit des estimateurs non biaisés de la proportion de survie attendue. Les estimateurs du rapport de la survie relative peuvent cependant être biaisés, car les proportions de survie observées le sont potentiellement. Cela est dû au fait que cette méthode ne prend pas en considération que les temps de suivi potentiels des patients sont de longueurs différentes.

Dans la méthode Ederer II, les temps de suivi hétérogènes observés et potentiels sont pris en compte. Par contre, la proportion de survie attendue est dépendante de la mortalité observée, ce qui implique des estimations biaisées du rapport de survie relative (Hakulinen, 1982). La mortalité observée pour un intervalle donné détermine quels sont les patients qui vont être à la base du calcul de p_{j2}^* dans le prochain intervalle. La proportion de survie attendue cumulative des patients est par conséquent dépendante de la létalité de la maladie en question pendant les intervalles précédents. En raison de

cela, la méthode Ederer II n'est pas recommandée pour l'estimation des proportions de survie attendue cumulatives, mais elle est tout de même une bonne estimation de la proportion de survie attendue des intervalles spécifiques.

Comme nous en avons déjà fait mention à la section 2.2.3, la méthode d'Hakulinen, contrairement aux deux autres méthodes, tient compte du nombre attendu de cas censurés pour chacun des intervalles de temps ainsi que de l'évolution de la cohorte étudiée. D'autre part, le calcul de la survie attendue tel que proposé dans cette méthode corrige les deux biais que comportent les méthodes Ederer I et II. Dans la méthode Ederer I, le calcul de la survie attendue ne tient pas compte des décès et des censures observées, alors que, dans la méthode Ederer II, la survie attendue dépend de la survie nette, c'est-à-dire de la maladie étudiée.

Lorsqu'on compare les estimations de la survie attendue à cinq ans obtenues à l'aide des trois méthodes proposées (tableaux 2.1, 2.2 et 2.3), on remarque qu'elles sont toutes très semblables. D'ailleurs, les trois méthodes proposées sont comparables pour des périodes de suivi allant jusqu'à 10 ans (Hakulinen, 1982). Par contre, la méthode d'Hakulinen est plus performante que les méthodes d'Ederer I et II pour des périodes de suivi plus longues.

Chapitre 3

Modélisation de la survie relative

Au chapitre précédent, nous avons étudié l'estimation de la survie relative. D'autres méthodes d'estimation, qui reposent cette fois sur la modélisation de l'équation 2.1, seront explorées dans le présent chapitre. Ces méthodes sont toutes basées sur une approche de maximisation de la vraisemblance pour l'estimation des paramètres du modèle.

3.1 Méthode d'Estève *et al.* (1990) : approche par maximisation de la vraisemblance

Une approche bien connue pour la modélisation de la survie relative, appelée approche par maximisation de la vraisemblance, a été proposée par Estève *et al.* (1990). Ce travail venait en fait parfaire les idées de Pocock *et al.* (1982), Buckley (1984) et Andersen *et al.* (1985). C'était la première fois que des formules explicites étaient diffusées pour l'implantation de la méthode. Comme son nom l'indique, cette méthode emploie la maximisation de la vraisemblance afin d'estimer les paramètres du modèle de survie relative et ce, à l'aide de données individuelles.

Soit $(t_i, a_i, \delta_i, \mathbf{x}_i)$, $i=1, \dots, n$, un échantillon de n individus où t_i est le temps de survie depuis le diagnostic pour le i ème sujet, a_i est l'âge au diagnostic du i ème sujet, δ_i est l'indicatrice de décès ($\delta_i = 1$ si t_i est le moment du décès et $\delta_i = 0$ si le temps de survie est censuré à t_i) et \mathbf{x}_i est le vecteur de covariables pour le i ème sujet. La fonction de vraisemblance, qui est associée au modèle de l'équation (2.1) et qui permet d'estimer

β , est donc donnée par

$$L(\beta) = \prod_{i=1}^n \exp\left(-\int_0^{t_i} \lambda(s; \mathbf{x}_i; a_i) ds\right) [\lambda(t_i; \mathbf{x}_i; a_i)]^{\delta_i},$$

où $\lambda(s; \mathbf{x}_i; a_i)$ est défini par l'équation (2.3).

La log-vraisemblance s'écrit donc :

$$l(\beta) = -\sum_{i=1}^n \left\{ \sum_{l=0}^{[t_i-1]} \lambda^*(a_i + l; \mathbf{x}_i) + (t_i - 1 - [t_i - 1])\lambda^*(a_i + [t_i]; \mathbf{x}_i) + t_i \exp(\mathbf{x}_i \beta) - \delta_i \ln[\lambda^*(a_i + [t_i]; \mathbf{x}_i) + \exp(\mathbf{x}_i \beta)] \right\}.$$

Il est à noter que si t_i est plus petit que 1, le premier terme de l'expression est nul.

La contribution d'un seul individu à la log-vraisemblance est

$$l_i(\beta) = -\left\{ t_i \exp(\mathbf{x}_i \beta) - \delta_i \ln[\lambda^*(a_i + [t_i]; \mathbf{x}_i) + \exp(\mathbf{x}_i \beta)] \right\}, \quad (3.1)$$

car $\sum_{l=0}^{[t_i-1]} \lambda^*(a_i + l; \mathbf{x}_i)$ et $(t_i - 1 - [t_i - 1])\lambda^*(a_i + [t_i]; \mathbf{x}_i)$ ne dépendent pas de β .

Le principal avantage lié à l'utilisation de cette méthode est qu'elle permet d'éviter les problèmes potentiels liés au regroupement de patients hétérogènes (du point de vue de la survie). La méthode d'Estève présente, par contre, quelques inconvénients. En effet, aucun moyen de vérifier l'ajustement du modèle n'aurait encore été proposé (Dickman *et al.*, 2004). Il ne serait pas non plus possible, selon Dickman *et al.* (2004), d'établir des diagnostics de régression, ni d'estimer les termes d'interactions. Enfin, la théorie entourant ce modèle ne permet pas encore de modéliser les covariables variant dans le temps (principalement en raison de limites informatiques), moyen utilisé pour contrôler les risques non proportionnels.

3.2 Étude d'un cas particulier du modèle d'Estève et *al.*

Considérons maintenant un cas particulier du modèle d'Estève et *al.*, où l'âge au diagnostic ne fait pas partie du vecteur de covariables définissant le i ème sujet. En d'autres mots, \mathbf{x}_i ne dépend pas de a_i . Nous allons en fait supposer que \mathbf{x}_i comporte K catégories distinctes, ce qui implique que nous disposons d'une table de mortalité pour chacune des strates.

Nous noterons

- $k = 1, \dots, K$ le nombre de strates, c'est-à-dire le nombre de valeurs possibles du vecteur \mathbf{x}_i ;
- $i = 1, \dots, n_k$ le nombre de sujets dans chacune des strates ;
- $j = 0, \dots, J$ les années d'une table de survie ;
- $\lambda_k^*(j)$ le taux de mortalité attendue pour l'âge j de la strate k . Notons que dans (3.1), $\lambda^*(a_i + [t_i]; \mathbf{x}_i) = \lambda_k^*(j)$ si $a_i + [t_i] = j$ et si \mathbf{x}_i correspond au vecteur de variables explicatives de la strate k ;
- $\delta_{jik} = \begin{cases} 1 & \text{si l'individu } i \text{ de la strate } k \text{ est décédé à l'âge } j \\ 0 & \text{sinon;} \end{cases}$
- $y_{jik} \in [0, 1]$ le temps à risque de l'individu i de la strate k à l'âge j , où $t_i = \sum_{j=0}^J y_{jik}$;
- $\delta_{jk} = \sum_{i=1}^{n_k} \delta_{jik}$ le nombre de décès qui surviennent à l'âge j de la strate k ;
- $y_{jk} = \sum_{i=1}^{n_k} y_{jik}$ le temps à risque total pour tous les individus à risque à l'âge j de la strate k ;
- $\delta_{jk}^* = y_{jk} \lambda_k^*(j)$ le nombre attendu de décès à l'âge j de la strate k (dû à d'autres causes qu'à la maladie d'intérêt et estimé à partir des taux de mortalité de la population générale).

Ainsi, l'approche proposée par Estève et *al.*, basée sur l'utilisation de données individuelles, peut être simplifiée si chacune des observations est visualisée en termes d'années de la table de survie. Plutôt que d'évaluer la log-vraisemblance pour chacun des sujets et de sommer pour tous les sujets, la log-vraisemblance est évaluée pour chacune des années de la table de survie, et ce, dans chacune des strates. En d'autres mots, l'échantillon $(t_i, a_i, \delta_i, \mathbf{x}_i)$ est réorganisé de façon à ce que nous ayons le temps à

risque total pour tous les individus, ainsi que le nombre de décès à chaque âge j de la table de survie pour la strate k . De ce fait, les années de la table de survie pour un seul individu représentent son expérience de survie et incluent les variables représentant le temps à risque (y_{jik}), l'indicateur de décès (δ_{jik}), le risque attendu ($\lambda_k^*(j)$) et des variables indicatrices pour chacune des composantes de β . De plus, nous notons que tous les individus de la strate k ont un \mathbf{x}_i égal à \mathbf{x}_k . À titre d'exemple, considérons l'individu i faisant partie de la strate k qui est décédé en 1992 à 76 ans, soit 3.25 ans après le diagnostic de la maladie ($t_i = 3.25, \delta_i = 1$). La table de mortalité pour cet individu est donc présentée à la table 3.1. Les $\lambda_k^*(j)$ présentés dans cette table sont tirés des tables québécoises de mortalité de 1986 et 1991 présentées en annexe.

TAB. 3.1 – Table de mortalité pour l'individu i de la strate k , décédé en 1992 à 76 ans, 3.25 ans après le diagnostic ($t_i = 3.25, \delta_i = 1$).

j	Années de suivi	Temps à risque (y_{jik})	δ_{jik}	$\lambda_k^*(j)^1$
0	1916	0	0	nd
\vdots	\vdots	\vdots	\vdots	\vdots
72	1988	0	0	0.04996
73	1989	1	0	0.04705
74	1990	1	0	0.05135
75	1991	1	0	0.05596
76	1992	0.25	1	0.06096
77	1993	0	0	0.06642
\vdots	\vdots	\vdots	\vdots	\vdots
100	2016	0	0	nd
Total sur j		3.25	1	

¹ nd = non déterminé, car les tables de mortalité ne sont pas disponibles

Si nous réécrivons maintenant la log-vraisemblance de l'équation (3.1) en réorganisant les données de cette façon, la contribution d'un individu de la strate k est

$$l_i(\beta) = \sum_{j=0}^J \left\{ \delta_{jik} \ln [\lambda_k^*(j) + \exp(\mathbf{x}_k \beta)] - y_{jik} \exp(\mathbf{x}_k \beta) \right\}.$$

Contrairement à ce que nous avons en (3.1), la base n'est pas ici l'individu, mais bien l'âge associé à la table de survie.

La contribution de tous les individus à la fonction log-vraisemblance se traduit alors ainsi

$$l(\beta) = \sum_i l_i(\beta) = \sum_{j=0}^J \sum_{k=1}^K \left\{ \delta_{jk} \ln[\lambda_k^*(j) + \exp(\mathbf{x}_k \beta)] - y_{jk} \exp(\mathbf{x}_k \beta) \right\}. \quad (3.2)$$

Or, cette log-vraisemblance est identique à celle d'un modèle Poisson. En effet, supposons que δ_{jk} suit une loi de Poisson de paramètre $\mu_{jk} = y_{jk}[\lambda_k^*(j) + \exp(\mathbf{x}_k \beta)]$. La log-vraisemblance pour β s'écrit alors (Breslow et Day, 1987), à une constante près,

$$l(\beta) = \sum_j \sum_k \left\{ \delta_{jk} \ln[\lambda_k^*(j) + \exp(\mathbf{x}_k \beta)] - y_{jk} [\lambda_k^*(j) + \exp(\mathbf{x}_k \beta)] \right\}, \quad (3.3)$$

où δ_{jk} est le nombre de décès qui surviennent à l'âge j de la strate k et y_{jk} est le temps à risque total pour tous les individus à risque à l'âge j de la strate k .

Par conséquent, l'équation (2.3) peut s'écrire

$$\ln(\mu_{jk} - \delta_{jk}^*) = \ln(y_{jk}) + \mathbf{x}_k \beta, \quad (3.4)$$

car $\lambda_k^*(j) = \delta_{jk}^*/y_{jk}$, où δ_{jk}^* est le nombre attendu de décès.

L'équation (3.4) est associée à un modèle linéaire généralisé avec une erreur de structure Poisson, un offset $\ln(y_{jk})$ et un lien logarithmique non standard défini par $\ln(\mu_{jk} - \delta_{jk}^*)$ (un lien standard aurait été identifié par $\ln(\mu_{jk})$). Nous parlons ici de lien logarithmique non standard, car la fonction de lien décrit comment $\mu_{jk} - \delta_{jk}^*$ et non μ_{jk} lie les variables explicatives au prédicteur.

3.3 Extension du modèle d'Estève et *al.* : modélisation Poisson

Jusqu'ici, l'estimation du modèle de survie relative (2.3) a employé des données individuelles et des données agrégées. Dans le cas des données agrégées, chaque ligne de la base de données correspondait à une année d'une table de survie. Par extension, le modèle de survie relative peut aussi être estimé à partir de données divisées en J intervalles de suivi qui peuvent être des années, des mois ou des jours (Dickman *et al.*, 2004). Plus explicitement, le temps à risque d'un individu (y_i) est réparti sur les j intervalles de suivi définis au début de l'étude. Nous notons

- t_j , $0 \leq t_j - t_{j-1} \leq 1$, pour $j = 0, 1, \dots, J$ les bornes des intervalles de suivi avec $t_0 = 0$;
- $a_{jik} = a_i + t_{j-1}$ l'âge de l'individu i de la strate k pour l'intervalle j ;
- n_{jk} le nombre d'individus en vie au début de l'intervalle j (à t_{j-1}) de la strate k ;
- y_{jik} le temps à risque de l'individu i de la strate k pour l'intervalle j (y_{jik} doit toujours être exprimé en termes d'années, car les tables de mortalité ne sont disponibles que sur une base annuelle);
- $\delta_{jik} = \begin{cases} 1 & \text{si l'individu } i \text{ de la strate } k \text{ est décédé pendant l'intervalle } j \\ 0 & \text{sinon;} \end{cases}$
- $\lambda_k^*(a_{ji})$ le taux de mortalité annuel attendu pour l'individu i de la strate k à l'âge a_{jik} ;
- $p_{jk}^*(i) = \begin{cases} [1 - \lambda_k^*(a_{ji})]^{(t_j - t_{j-1})} & \text{si } y_{jik} > 0 \\ 0 & \text{sinon} \end{cases}$
la probabilité de survie attendue pour l'individu i de la strate k pour l'intervalle de suivi j (voir section 1.3);
- la probabilité de survie attendue pour l'intervalle j de la strate k

$$p_{jk}^* = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} p_{jk}^*(i). \quad (3.5)$$

Cette expression rejoint la définition du calcul de la survie attendue de la méthode Ederer II (section 2.2.2).

Pour illustrer ceci, reprenons l'exemple de la section 3.2 pour six intervalles de suivi de longueurs différentes ($t_0 = 0$, $t_1 = 6$ mois, $t_2 = 1$ an, $t_3 = 2$ ans, $t_4 = 3$ ans, $t_5 = 4$ ans et $t_6 = 5$ ans), où nous considérons un individu décédé en 1992 à 76 ans, c'est-à-dire 3.25 ans après le diagnostic de la maladie (voir table 3.2).

TAB. 3.2 – Intervalles de suivi pour l'individu i de la strate k , décédé en 1992 à 76 ans, 3.25 ans après le diagnostic.

j	Intervalle	a_{jik}	Temps à risque (y_{jik})	δ_{jik}	$\lambda_k^*(a_{ji})$	$p_{jk}^*(i)$
1	$[0, t_1]$	73	0.5	0	0.04705	0.97619
2	$[t_1, t_2]$	73	0.5	0	0.04705	0.97619
3	$[t_2, t_3]$	74	1	0	0.05135	0.94865
4	$[t_3, t_4]$	75	1	0	0.05596	0.94404
5	$[t_4, t_5]$	76	0.25	1	0.06096	0.98440
6	$[t_5, t_6]$	77	0	0	0.06096	0

Ce qu'il y a de particulier avec cette méthode, c'est que les $\lambda_k^*(j)$ ne sont pas disponibles lorsque les strates sont associées au suivi du patient. Le risque attendu pour un individu est donc déterminé à l'aide de l'âge atteint à chacun des intervalles de suivi. Une telle réorganisation des données implique que chacun des intervalles de suivi hérite des covariables de l'observation originale, comme par exemple l'âge au diagnostic, le sexe ou même le site de la tumeur dans le cas des cancers.

La contribution d'un individu de la strate k à la log-vraisemblance est définie par :

$$l_i(\beta) = \sum_{j=1}^J \left\{ \delta_{jik} \ln [\lambda_k^*(a_{ji}) + \exp(\mathbf{x}_k \beta)] - y_{jik} \exp(\mathbf{x}_k \beta) \right\}.$$

Et la contribution de tous les individus est

$$l(\beta) = \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{n_k} \left\{ \delta_{jik} \ln [\lambda_k^*(a_{ji}) + \exp(\mathbf{x}_k \beta)] - y_{jik} \exp(\mathbf{x}_k \beta) \right\}. \quad (3.6)$$

Puisque $\lambda_k^*(a_{ji})$ varie d'un individu à l'autre dans l'équation (3.6), il n'est pas possible d'évaluer de façon simple la somme sur i comme dans l'équation (3.2). Une alternative à la maximisation de la vraisemblance peut donc être envisagée. En effet, il est possible de travailler sur des approximations à partir d'une pseudo-vraisemblance.

À cet effet, nous considérons δ_{jik} , une variable aléatoire de distribution Bernouilli pour laquelle

$$P(\delta_{jik} = 1) = 1 - \exp\left\{-\int_{t_{j-1}}^{t_j} [\lambda_k^*(a_{ji}) + \exp(\mathbf{x}_k\beta)] dt\right\}$$

$$\approx (t_j - t_{j-1})[\lambda_k^*(a_{ji}) + \exp(\mathbf{x}_k\beta)].$$

De plus, $\delta_{jk} = \sum_{i=1}^{n_{jk}} \delta_{jik}$ est, pour l'intervalle j de la strate k , une somme de variables aléatoires de Bernouilli avec probabilités de succès variables.

Ainsi,

$$E(\delta_{jk}) \cong (t_j - t_{j-1}) \left\{ \sum_{i=1}^{n_{jk}} \lambda_k^*(a_{ji}) + n_{jk} \exp(\mathbf{x}_k\beta) \right\}$$

$$\cong \delta_{jk}^* + (t_j - t_{j-1}) n_{jk} \exp(\mathbf{x}_k\beta), \quad (3.7)$$

où $(t_j - t_{j-1}) \sum_{i=1}^{n_{jk}} \lambda_k^*(a_{ji})$ est associé à la mortalité naturelle, $(t_j - t_{j-1}) n_{jk} \exp(\mathbf{x}_k\beta)$ est associé à l'excès de mortalité et $\delta_{jk}^* = n_{jk}(1 - p_{jk}^*)$, où p_{jk}^* est définie par 3.5.

En fait, δ_{jk}^* correspond à la partie de l'espérance d'une loi binomiale attribuable à la mortalité dans la population générale. Il s'agit d'une loi binomiale de paramètres $(n_{jk}, 1 - p_{jk})$, où p_{jk} est la proportion de survie observée des n_{jk} sujets de la strate k qui étaient à risque pendant l'intervalle j . En d'autres mots, nous posons l'hypothèse que les δ_{jk} suivent une loi binomiale de paramètres n_{jk} et $1 - p_{jk}$. C'est en fait l'hypothèse qu'ont posé [Hakulinen et Tenkanen \(1987\)](#) pour modéliser la survie relative. Leur méthode est présentée à la section suivante.

Il est aussi possible d'estimer δ_{jk}^* à l'aide d'une méthode similaire à celle de la section précédente, où $\lambda_k^*(j) = \delta_{jk}^*/y_{jk}$. Une approximation en termes de taux est alors

considérée, où δ_{jk}^* s'écrit

$$\delta_{jk}^* = -\frac{\ln(p_{jk}^*)}{(t_j - t_{j-1})} y_{jk}. \quad (3.8)$$

Or, nous pouvons supposer, comme à la section précédente, que les δ_{jk} suivent une loi de Poisson, mais cette fois de paramètre $\mu_{jk} = \delta_{jk}^* + (t_j - t_{j-1})n_{jk}\exp(\mathbf{x}_k\beta)$, où les δ_{jk}^* sont définis par 3.8.

Le modèle peut donc être estimé, tout comme à la section 3.2, par un modèle linéaire généralisé d'erreur poisson avec offset $\ln[(t_j - t_{j-1})n_{jk}]$ et avec un lien logarithmique non standard $\ln(\mu_{jk} - \delta_{jk}^*)$. Ainsi,

$$\ln \left[\frac{\mu_{jk} - \delta_{jk}^*}{(t_j - t_{j-1})n_{jk}} \right] = \mathbf{x}_k\beta. \quad (3.9)$$

3.3.1 Extension du modèle d'Estève et al. : modélisation binomiale

Les méthodes d'estimation de la survie relative, étudiées dans les premières sections de ce chapitre, reposent sur la modélisation du risque. Le modèle régressif proposé par [Hakulinen et Tenkanen \(1987\)](#) repose, quant à lui, sur la modélisation de la survie relative. Cette méthode permet donc une modélisation plus directe de la survie. Elle nécessite d'ailleurs moins d'approximations que les méthodes précédentes.

Supposons que le nombre de décès δ_{jk} pour l'intervalle j de la strate k est une variable aléatoire de loi binomiale

$$\delta_{jk} \sim \text{Bin}(n_{jk}, 1 - p_{jk}),$$

où p_{jk} est la proportion de survie observée d'un intervalle spécifique, pour la strate k et l'intervalle j .

Cette formulation montre que le modèle de survie relative peut s'écrire comme un modèle linéaire généralisé, caractérisé, pour une strate et un intervalle de temps spécifique, par une erreur de structure binomiale et une fonction de lien log-logarithmique non standard $\ln \left[-\ln \frac{p_{jk}}{p_{jk}^*} \right]$. Ainsi,

$$\ln \left[-\ln \frac{p_{jk}}{p_{jk}^*} \right] = \mathbf{x}_k \beta, \quad (3.10)$$

où p_{jk}^* est la proportion de survie attendue d'un intervalle spécifique, pour la strate k et l'intervalle j . L'équation 3.5 définit p_{jk}^* .

Cette méthode implique que le taux de survie par strate et intervalle de temps est la moyenne des probabilités de survie attendue pour les individus vivants au début de l'intervalle. Ce mode de calcul s'avère être un handicap lorsque les intervalles de temps sont longs ou lorsque la population est hétérogène. Aussi, l'utilisation des strates entraînerait, selon [Giorgi \(2002\)](#), un manque de puissance et des estimations moins précises en cas de petits effectifs ou en cas de survie à très long terme. Ce modèle permet, par contre, de calculer les pertes d'espérance de vie ([Foucher et al., 1994](#)).

Chapitre 4

Application aux accidents vasculaires cérébraux

Le but de ce chapitre est d'illustrer et de comparer les différentes méthodes d'estimation de la survie relative décrites aux chapitres précédents. Pour ce faire, une banque de données, issue d'une étude sur les accidents vasculaires cérébraux (AVC), a été utilisée.

4.1 Mise en situation

L'accident vasculaire cérébral (AVC) est l'une des plus importantes causes de décès au Canada ([Fondation des maladies du coeur du Canada, 1999](#)). Il est aussi la deuxième cause la plus fréquente d'invalidité d'origine neurologique, après la maladie d'Alzheimer, dans les pays occidentaux ([Merck & Co., 2004](#)). Même si son incidence a diminué au cours des dernières années, il reste la principale cause de soins en établissement spécialisé en raison de la perte d'indépendance chez l'adulte.

L'accident vasculaire cérébral est un trouble vasculaire cérébral touchant les vaisseaux sanguins qui amènent le sang au cerveau. En d'autres mots, il s'agit d'un arrêt subit du fonctionnement du cerveau. Cet arrêt peut être causé par l'interruption de la circulation sanguine vers le cerveau (AVC de type ischémique) ou par la rupture d'un vaisseau sanguin du cerveau (AVC de type hémorragique). L'interruption de la

circulation sanguine ou la rupture d'un vaisseau sanguin implique qu'une partie du cerveau est donc privée de l'apport sanguin nécessaire, ce qui provoque la mort des cellules cérébrales (neurones) de la partie du cerveau atteinte. Étant donné que les cellules nerveuses sont touchées, les parties du corps qui en dépendent ne peuvent plus fonctionner. Les conséquences peuvent être très dévastatrices et souvent même permanentes. Les AVC les plus fréquents sont la thrombose et l'embolie cérébrale, qui sont tous deux des AVC de type ischémique. Ils représentent à eux seuls de 70 à 80 pour cent de tous les AVC.

La cause majeure de l'AVC demeure l'athérosclérose, dont les facteurs de risque les plus importants sont l'hypercholestérolémie, l'hypertension artérielle et le tabagisme ([Service Vie, 2003](#)). L'obésité, le diabète sucré, l'hypertriglycéridémie, la sédentarité et le stress sont aussi des facteurs influents. L'âge et le sexe sont également deux importants facteurs de risque pour les AVC. En effet, le risque de subir un AVC augmente considérablement avec l'âge et les femmes sont plus souvent touchées que les hommes. Presque 60 % des AVC qui surviennent chaque année au Canada touchent les femmes ([Fondation des maladies du coeur du Canada, 1999](#)).

4.2 Description de la base de données

Pour la création de cette base de données, l'identification des cas rejoint tous les résidents québécois (peu importe leur âge) qui ont été hospitalisés au Québec pour un accident vasculaire cérébral (CIM-9¹, codes 430-438). Le fichier MED-ECHO (Maintenance et Exploitation des Données pour l'Étude de la Clientèle HOspitalière) a donc été utilisé pour connaître les hospitalisations pour cette maladie. Ce fichier a par la suite été apparié au fichier de la Régie de l'Assurance-Maladie du Québec (RAMQ), ainsi qu'au fichier démographique des décès du Ministère de la Santé et des Services Sociaux du Québec pour compléter les informations sur le statut vital, le sexe et la date de naissance des individus. Pour tous les sujets de ces trois fichiers, les variables disponibles sont listées à l'annexe B.

Un des objectifs associé à cette base de données était d'estimer la survie relative aux AVC. Il s'agit d'une première au Québec, et probablement même au Canada. Jusqu'à ce jour, la survie relative a plutôt été utilisée dans les études sur le cancer ([Louchini](#)

¹Classification Internationale des Maladies, version 9 : 1979 à 1999

et Beaupré, 2003; Louchini, 2002; Ellison *et al.*, 2001). La survie relative est pourtant la méthode toute désignée pour étudier la survie des individus ayant subi un accident vasculaire cérébral. En effet, les données disponibles sont de type populationnel et il est souvent impossible de déterminer si la cause exacte du décès est l'AVC ou s'il s'agit d'une autre cause.

Pour estimer ou modéliser la survie relative, seulement les variables suivantes ont été utilisées : l'année de l'hospitalisation, la date d'admission à l'hôpital, la date de décès provenant du fichier de décès, la date de décès provenant du fichier de la RAMQ, le type de décès, la date de naissance provenant du fichier MED-ECHO, la date de naissance provenant du fichier de décès, la date de naissance provenant du fichier de la RAMQ, la date de sortie de l'hôpital, le numéro identifiant personnel, le sexe provenant du fichier MED-ECHO, le sexe provenant du fichier de décès et le sexe provenant du fichier de la RAMQ. Un astérisque annote ces variables à l'annexe B.

Par ailleurs, chaque enregistrement du fichier MED-ECHO correspond à une hospitalisation. Par conséquent, plusieurs hospitalisations peuvent être enregistrées pour un même sujet. Étant donné que l'étude porte essentiellement sur la première hospitalisation pour maladies vasculaires cérébrales, seule cette dernière a été sélectionnée pour les analyses. Nous avons aussi restreint l'analyse aux individus de 25 ans et plus lors de l'hospitalisation (le nombre de cas est très faible chez les moins de 25 ans) et pour lesquels l'accident vasculaire cérébral est survenu entre 1990 et 1992.

Avant de passer à l'exploration des données et aux analyses, plusieurs manipulations de la base de données ont dû être effectuées. Tout d'abord, les informations concernant le sexe, la date de naissance et la date de décès ont dû être complétées à l'aide du fichier de la RAMQ et du fichier sur les décès. Dans le cas du sexe, nous avons privilégié la valeur provenant du fichier de la RAMQ. Lorsque cette valeur était manquante, la donnée était récupérée du fichier MED-ECHO. Une démarche semblable a aussi permis de compléter les informations concernant la date de naissance, où l'ordre de priorité des fichiers était : RAMQ, MED-ECHO, fichier de décès.

En ce qui concerne la date de décès, nous avons d'abord favorisé la date provenant du fichier de décès et ensuite celle provenant du fichier de la RAMQ. Comme il s'agit d'un suivi passif des individus basé sur l'utilisation de fichiers administratifs, nous voulions récupérer des trois fichiers disponibles le plus de décès possible, entre 1990 et 1992.

Nous avons donc attribué, comme date de décès, la date de sortie de l'hôpital aux sujets qui sont décédés à l'hôpital. Lorsque l'information sur le décès d'un individu n'était pas trouvée dans l'un de ces trois fichiers, nous avons supposé que cette personne était toujours vivante et résidait toujours au Québec. Nous avons donc fait l'hypothèse que la banque de données ne contenait pas de données censurées.

La variable survtime, le temps de survie depuis le diagnostic, a ensuite été créée. Elle est définie comme étant la différence entre la date de décès et la date d'admission à l'hôpital. Une variable pour l'âge des individus lors de l'hospitalisation a également été créée (différence entre la date de l'hospitalisation et la date de naissance). Cette variable a ensuite été catégorisée de la façon suivante : 25 à 44 ans, 45 à 54 ans, 55 à 64 ans, 65 à 69 ans, 70 à 74 ans, 75 à 79 ans, 80 à 84 ans et 85 ans et plus.

Dans un autre ordre d'idées, soulignons que cinq intervalles de suivi spécifiques, qui suivent l'hospitalisation pour un AVC, ont préalablement été déterminés, à savoir de 0 à 7 jours, de 7 à 28 jours, de 28 à 60 jours, de 60 jours à 1 an (ou 365.25 jours) et de 1 an à 2 ans (ou 730.50 jours). Les premiers intervalles choisis sont très courts, car les premiers jours suivant l'AVC sont souvent critiques. Les deux derniers intervalles ont, quant à eux, été choisis pour étudier la survie à long terme après l'hospitalisation.

4.3 Exploration des données

Il apparaît, dans un premier temps, essentiel d'établir le nombre de cas d'AVC qui sont survenus entre 1990 et 1992, ainsi que le nombre de décès survenus parmi les sujets qui ont été hospitalisés pour un AVC. Comme le résume le tableau 4.1, la base de données contient 34 143 sujets, 17 227 hommes et 16 916 femmes, dont 11 689 sont décédés (5722 décès chez les hommes et 5967 décès chez les femmes), d'un AVC ou d'une autre cause, deux ans ou moins après l'hospitalisation.

Nous remarquons de plus, que la proportion de décès de toutes causes, peu importe le sexe, a diminué légèrement entre 1990 et 1992. Nous pouvons aussi déduire de ces résultats, que 34.2% des sujets ayant subi un AVC entre 1990 et 1992 sont décédés dans les deux années qui ont suivi l'hospitalisation. De plus, nous constatons que le nombre de décès est toujours supérieur chez les femmes. Ce résultat vient, en fait, corroborer celui de Statistique Canada ([Fondation des maladies du coeur du Canada, 1999](#), p.72)

TAB. 4.1 – Nombre de cas d’AVC et nombre de décès de toutes causes parmi ces cas survenus dans la période de deux ans suivant l’hospitalisation, pour les années 1990 à 1992.

Années	Hommes			Femmes			Total		
	Cas d’AVC	Décès	%	Cas d’AVC	Décès	%	Cas d’AVC	Décès	%
1990	5681	1940	34.1	5394	1952	36.2	11075	3892	35.1
1991	5797	1951	33.7	5655	1980	35.0	11452	3931	34.3
1992	5749	1831	31.8	5867	2035	34.7	11616	3866	33.3
Total	17 227	5722	33.2	16 916	5967	35.3	34 143	11 689	34.2

qui a montré que depuis 1950, le nombre de décès attribuables aux maladies vasculaires cérébrales est demeuré plus élevé chez les femmes que chez les hommes. Nous pouvons aussi observer dans le tableau 4.1 que la proportion de décès est toujours supérieure chez les femmes.

Par contre, la proportion de décès dans les deux années suivant l’hospitalisation, pour des hospitalisations survenues entre 1990 et 1992, est plus élevée chez les hommes dans toutes les catégories d’âges, sauf chez les 45 à 54 ans. Le tableau 4.2 présente, à cet effet, la répartition du nombre de sujets qui ont subi un AVC, du nombre de décès dans les deux ans suivant le diagnostic et de la proportion de décès pour les huit catégories d’âge lors de l’hospitalisation. En outre, les hommes de cette population sont en moyenne plus jeunes que les femmes (moyenne d’âge de 68.5 ans chez les hommes et de 72.4 ans chez les femmes).

TAB. 4.2 – Nombre de cas d’AVC et nombre de décès de toutes causes parmi ces cas survenus dans la période de deux ans suivant l’hospitalisation, pour chacune des catégories d’âge.

Âge à l’hospitalisation	Hommes			Femmes			Total		
	Cas d’AVC	Décès	Prop.(%)	Cas d’AVC	Décès	Prop.(%)	Cas d’AVC	Décès	Prop.(%)
25-44	723	125	17.3	725	95	13.1	1448	220	15.2
45-54	1423	218	15.3	963	167	17.3	2386	385	16.1
55-64	3509	701	20.0	2058	390	19.0	5567	1091	19.6
65-69	2842	728	25.6	2026	461	22.8	4868	1189	24.4
70-74	2941	989	33.6	2588	785	30.3	5529	1774	32.1
75-79	2692	1119	41.6	3032	1124	37.1	5724	2243	39.2
80-84	1858	1041	56.0	2911	1348	46.3	4769	2389	50.1
85 et +	1239	801	64.6	2613	1597	61.1	3852	2398	62.3
Total	17 227	5722	33.2	16 916	5967	35.3	34 143	11 689	34.2

4.4 Estimation de la survie relative

Comme nous l'avons mentionné à plusieurs reprises dans les chapitres précédents, la survie relative est définie comme le rapport entre la survie observée et la survie attendue. La section 4.4.1 présente la méthode choisie pour le calcul de la survie observée. La section 4.4.2 présente, quant à elle, une comparaison de deux des trois méthodes d'estimation de la survie attendue, alors que la section 4.4.3 expose les résultats de l'estimation du ratio de survie relative.

4.4.1 Survie observée

Dans la base de données sur les accidents vasculaires cérébraux, on suppose que tout individu pour lequel on ne retrouve pas de date de décès est toujours vivant et réside toujours au Québec. On émet donc l'hypothèse que les données n'incluent pas de temps de survie censurés. Par conséquent, les méthodes de Kaplan-Meier et actuarielle, pour le calcul de la survie observée, devraient donner les mêmes estimations. La méthode actuarielle a tout de même été préférée à celle de Kaplan-Meier, notamment parce qu'elle permet l'estimation de la survie observée pour des données groupées (voir section 2.1.4), et que ce sont des données groupées qui seront utilisées pour les AVC.

La survie observée par la méthode actuarielle a été calculée pour chacun des intervalles de suivi définis plus tôt, conditionnellement à la survie de l'intervalle précédent, à l'aide de l'équation 2.4. Elle est présentée, par groupe d'âge et par sexe, au tableau 4.3. Il n'est pas surprenant d'observer que la survie aux AVC varie beaucoup d'une catégorie d'âge à une autre. D'ailleurs, un test du log-rank a montré que les probabilités de survie observées diffèrent significativement (p -value < 0.0001) d'une catégorie d'âge à l'autre. La survie a tendance à être supérieure chez les sujets les plus jeunes. Par contre, la probabilité de survie chez les 65-69 ans est supérieure à celle des autres catégories d'âge à sept jours. Autre fait intéressant, la survie observée à sept jours est inférieure à celle des autres périodes de suivi chez les 25-64 ans.

Il est aussi important de noter que la probabilité de survie à sept jours chez les femmes est toujours inférieure à celle chez les hommes et ce, peu importe la catégorie d'âge, alors que la probabilité de survie à long terme, c'est-à-dire à un an et à deux ans, est toujours supérieure chez les femmes.

TAB. 4.3 – Survie observée (%) spécifique aux intervalles (Int) et cumulative (Cum), par groupe d'âge et par sexe.

	Groupes d'âge															
	25-44		45-54		55-64		65-69		70-74		75-79		80-84		85 et +	
	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum
Hommes																
7 jours	91.4	91.4	94.0	94.0	95.3	95.3	95.3	95.3	94.5	94.5	93.3	93.3	91.9	91.9	89.7	89.7
28 jours	96.4	88.1	97.5	91.6	96.1	91.7	96.2	91.6	95.1	89.8	93.2	87.0	89.5	82.2	86.1	77.2
60 jours	99.1	87.3	99.2	90.9	98.5	90.3	97.5	89.3	96.8	87.0	96.5	83.9	93.2	76.6	89.9	69.4
1 an	97.9	85.5	96.6	87.8	94.9	85.7	93.3	83.3	89.0	77.4	85.3	71.5	78.4	60.1	77.4	53.8
2 ans	97.7	83.5	97.8	85.9	95.5	81.8	92.9	77.4	89.9	69.6	87.7	62.8	82.1	49.3	74.9	40.3
Femmes																
7 jours	91.3	91.3	91.4	91.4	94.5	94.5	95.0	95.0	93.6	93.6	93.0	93.0	91.8	91.8	89.7	89.7
28 jours	98.2	89.7	97.0	88.7	96.9	91.5	95.8	91.0	95.4	89.3	92.9	86.3	90.7	83.3	85.2	76.5
60 jours	98.9	88.7	99.5	88.3	98.7	90.3	98.1	89.3	97.0	86.6	95.8	82.8	94.4	78.6	91.7	70.1
1 an	98.4	87.3	96.6	85.3	95.4	86.2	94.4	84.3	89.3	77.4	89.0	73.6	85.1	66.8	78.1	54.8
2 ans	99.7	87.0	97.8	83.4	96.2	82.9	94.4	79.5	93.3	72.2	90.1	66.3	86.8	58.0	79.2	43.4
Total																
7 jours	91.4	91.4	92.9	92.9	95.0	95.0	95.2	95.2	94.1	94.1	93.1	93.1	91.8	91.8	89.7	89.7
28 jours	97.3	88.9	97.3	90.4	96.4	91.6	96.0	91.4	95.2	89.6	93.0	86.6	90.2	82.9	85.5	76.7
60 jours	99.0	88.0	99.3	89.8	98.6	90.3	97.7	89.3	96.9	86.8	96.1	83.3	93.9	77.8	91.1	69.9
1 an	98.2	86.4	96.6	86.8	95.1	85.8	93.7	93.7	89.1	77.4	87.2	72.6	82.5	64.2	77.9	54.4
2 ans	98.7	85.3	97.8	84.9	95.8	82.2	93.5	83.7	91.5	70.8	89.0	64.7	85.1	54.6	77.8	42.4

4.4.2 Survie attendue

Le tableau 4.4 présente la survie attendue spécifique aux intervalles et cumulative pour les cinq intervalles de suivi, par sexe, pour l'ensemble des individus, et calculée à l'aide des méthodes Ederer I et Ederer II (voir sections 2.1.1 et 2.1.2).

TAB. 4.4 – Estimation de la survie attendue spécifique aux intervalles (Int) et cumulative (Cum), par sexe, à l'aide des méthodes Ederer I et Ederer II.

Intervalles*	Ederer I						Ederer II					
	H		F		Total		H		F		Total	
	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum
0 - 7	0.99904	0.99904	0.99919	0.99919	0.99911	0.99911	0.99904	0.99904	0.99919	0.99919	0.99911	0.99911
7 - 28	0.99615	0.99520	0.99677	0.99597	0.99646	0.99558	0.99716	0.99620	0.99762	0.99681	0.99739	0.99650
28 - 60	0.99177	0.98704	0.99308	0.98910	0.99242	0.98806	0.99582	0.99203	0.99655	0.99337	0.99618	0.99269
60 - 1 an	0.95145	0.93967	0.95897	0.94898	0.95517	0.94428	0.96233	0.95466	0.96887	0.96245	0.96552	0.95846
1 an - 2 ans	0.90399	0.85404	0.91803	0.87511	0.91095	0.86448	0.95802	0.91459	0.96536	0.92911	0.96159	0.92165

*Si non spécifiés, les intervalles sont en jours.

Nous remarquons que pour le premier intervalle, la survie attendue est la même pour les méthodes Ederer I et II. Par contre, la méthode Ederer I donne des estima-

tions légèrement inférieures à celles de la méthode Ederer II pour les intervalles de suivi au-delà de 7 jours. Hakulinen (1982) a d'ailleurs prouvé que dans le cas de trois types de cancer, les méthodes Ederer I, Ederer II et Hakulinen donnent des estimations très similaires pour des intervalles de suivi inférieurs à 10 ans. Nous aurions peut-être pu penser que ce ne serait pas le cas avec une maladie pour laquelle le temps de survie est habituellement plus court, comme l'AVC, mais les résultats semblent indiquer la même tendance que pour les cancers. De plus, nous pouvons constater que peu importe la méthode, la survie attendue chez les femmes est toujours supérieure à celle chez les hommes. Cela est entre autres dû au fait qu'à tout âge, la probabilité de survie d'une femme était supérieure à celle d'un homme au Québec en 1991 (voir les tables de mortalité à l'annexe A).

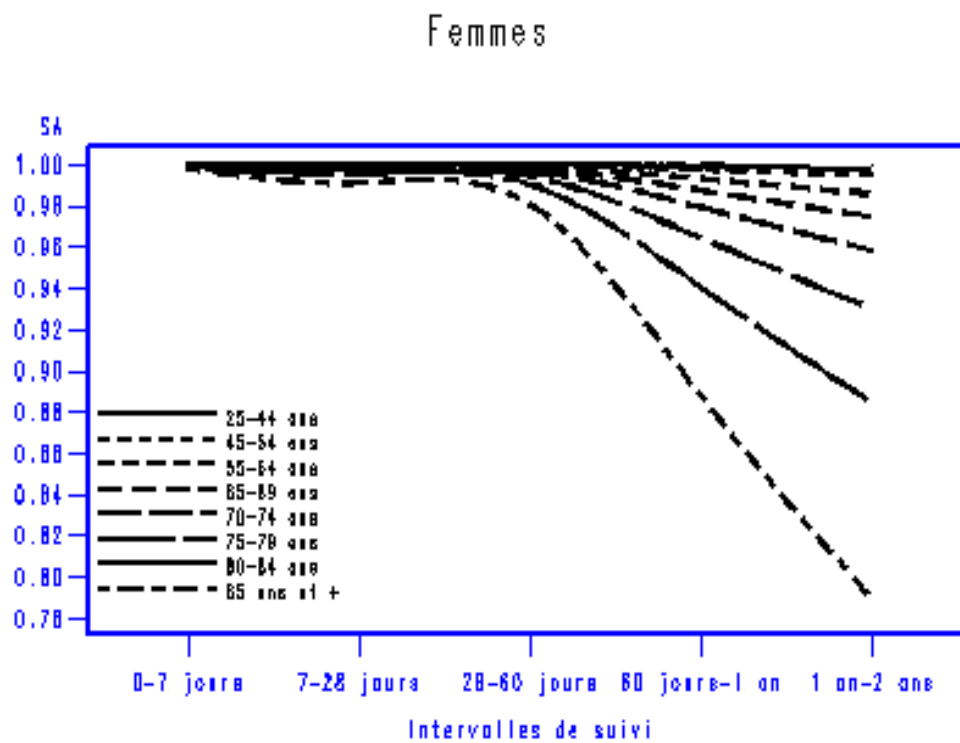
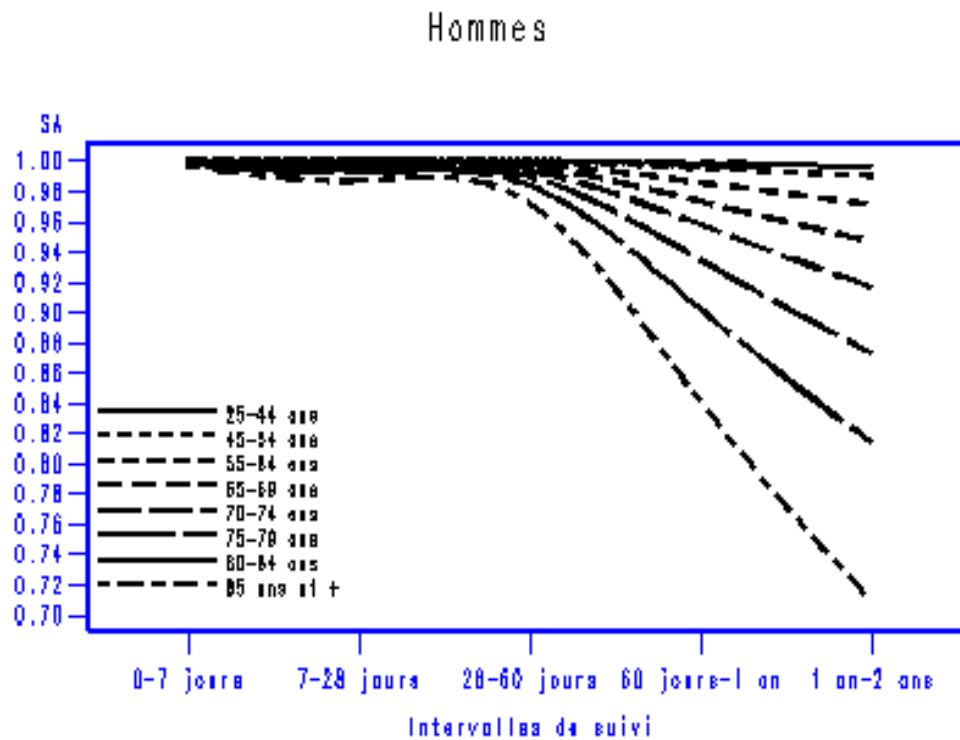
Mentionnons en outre, que la survie attendue ne dépend pas seulement du sexe, mais aussi de l'âge au diagnostic. En effet, l'analyse de la figure 4.1, qui présente la survie attendue calculée à partir de la méthode Ederer II pour chacune des catégories d'âge, par sexe, indique que si ces sujets n'avaient pas souffert d'un AVC, mais avaient appartenu à un groupe de la population générale comparable, leur survie attendue aurait été très dépendante de l'âge.

La méthode Ederer II a été choisie pour le calcul de la survie attendue, notamment car elle produit une meilleure estimation que la méthode Ederer I pour les proportions de survie attendue d'intervalles de suivi spécifiques. Comme on le voit au tableau 4.4, la méthode Ederer I, par rapport à la méthode Ederer II, sous-estime l'ensemble des estimations de la survie attendue. Cette sous-estimation vient du fait que cette méthode utilise l'information de l'ensemble des individus sur tous les intervalles, même si ces derniers ne sont pas sous suivi au début de tous les intervalles. Le choix de la méthode Ederer I résulterait donc en une surestimation du ratio de survie relative.

4.4.3 Survie relative

Le tableau 4.5 présente le ratio de survie relative spécifique aux intervalles de suivi et cumulative pour les sujets qui ont été hospitalisés pour un accident vasculaire cérébral entre 1990 et 1992. Les résultats présentés pour chacun des intervalles sont, en réalité, conditionnels au fait d'avoir survécu à l'intervalle de suivi précédent.

FIG. 4.1 – Survie attendue cumulative calculée à l'aide de la méthode Ederer II, par catégories d'âge pour chacun des intervalles de suivi.



TAB. 4.5 – Survie relative (%) spécifique aux intervalles (Int) et cumulative (Cum), par groupe d'âge et par sexe.

	Groupes d'âge															
	25-44		45-54		55-64		65-69		70-74		75-79		80-84		85 et +	
	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum	Int	Cum
Hommes																
7 jours	91.4	91.4	94.0	94.0	95.4	95.4	95.3	95.3	94.6	94.6	93.4	93.4	92.1	92.1	90.1	90.1
28 jours	96.4	88.1	97.6	91.7	96.2	91.8	96.3	91.8	95.3	90.1	93.6	87.4	90.1	82.9	86.9	78.3
60 jours	99.1	87.3	99.2	90.9	98.6	90.5	97.7	89.7	97.2	87.6	97.0	84.8	94.0	78.0	91.3	71.4
1 an	98.1	85.6	97.0	88.2	96.1	86.9	95.4	85.6	92.3	80.9	90.3	76.6	85.5	66.6	89.5	64.0
2 ans	97.9	83.9	98.3	86.8	97.0	84.3	95.6	81.8	94.0	76.0	93.9	72.0	91.0	60.6	88.9	56.9
Femmes																
7 jours	91.3	91.3	91.4	91.4	94.5	94.5	95.0	95.0	93.6	93.6	93.0	93.0	91.9	91.9	90.0	90.0
28 jours	98.2	89.7	97.1	88.7	96.9	91.5	95.9	91.1	95.5	89.4	93.1	86.6	91.0	83.7	85.8	77.2
60 jours	98.9	88.7	99.6	88.3	98.8	90.4	98.2	89.5	97.2	86.9	96.1	83.2	94.9	79.4	92.7	71.6
1 an	98.5	87.4	96.8	85.5	96.0	86.8	95.4	85.4	90.9	79.0	91.7	76.4	89.6	71.1	86.2	61.7
2 ans	99.8	87.2	98.1	83.8	96.9	84.1	95.6	81.6	95.3	75.3	93.5	71.4	92.3	65.6	89.0	54.9
Total																
7 jours	91.4	91.4	92.9	92.9	95.0	95.0	95.2	95.2	94.1	94.1	93.2	93.2	92.0	92.0	90.0	90.0
28 jours	97.3	88.9	97.4	90.5	96.5	91.7	96.1	91.5	95.4	89.8	93.3	87.0	90.7	83.4	86.2	77.6
60 jours	99.0	88.0	99.3	89.9	98.7	90.5	97.9	89.6	97.2	87.3	96.6	84.0	94.6	78.8	92.2	71.5
1 an	98.3	86.5	96.9	87.1	96.0	86.9	95.4	85.5	91.7	80.0	91.1	76.5	88.0	69.4	87.2	62.4
2 ans	98.9	85.5	98.2	85.6	97.0	84.2	95.6	81.7	94.6	75.7	93.7	71.6	91.8	63.7	89.0	55.5

Le ratio de survie relative cumulé peut être interprété comme la proportion de sujets en vie après un intervalle de suivi spécifique, dans la situation hypothétique où l'accident vasculaire cérébral serait la seule cause possible de décès. En d'autres mots, une survie relative inférieure à 100% signifie que, pendant l'intervalle spécifié, la mortalité des individus qui ont subi un AVC a été plus importante que la mortalité des individus comparables de la population générale, mais indemnes de la maladie (Foucher *et al.*, 1994). Cette interprétation n'est toutefois valide qu'en supposant que la mortalité non liée aux AVC est indépendante de la mortalité liée aux AVC. En général, il apparaît que la proportion des sujets en vie, après un intervalle, diminue avec l'âge et diminue lorsque le temps augmente après l'hospitalisation.

4.5 Modélisation de la survie relative

Comme nous l'avons vu au chapitre III, l'équation de base permettant de modéliser la survie relative (équation (2.1)) peut être estimée à l'aide de plusieurs approches. Le but de cette section est de comparer certaines de ces approches pour ensuite établir laquelle serait la meilleure pour l'application aux accidents vasculaires cérébraux.

Cinq covariables ont été choisies pour expliquer la survie relative. Il s'agit du sexe, de l'âge au moment de l'hospitalisation, de l'année de l'hospitalisation, du temps de suivi et du type d'accident vasculaire cérébral. En fait, le temps de suivi apparaît dans la matrice des covariables sous forme d'intervalles. Voici la justification du choix de chacune de ces variables :

Sexe (SEX) : Le sexe est un facteur de risque important, associé à la naissance des accidents vasculaires cérébraux. En effet, le taux de mortalité ajusté par AVC est supérieur chez les hommes au Québec, alors que le nombre de cas annuel est supérieur chez les femmes (Bouchard et Louchini, 2001).

Âge lors de l'hospitalisation (AGE_DIAG) : L'âge est le facteur de risque dominant pour toutes les maladies cérébro-vasculaires, incluant les AVC (Fondation des maladies du coeur du Canada, 1999). Le risque de subir un AVC ainsi que le taux de mortalité par AVC augmentent avec l'âge.

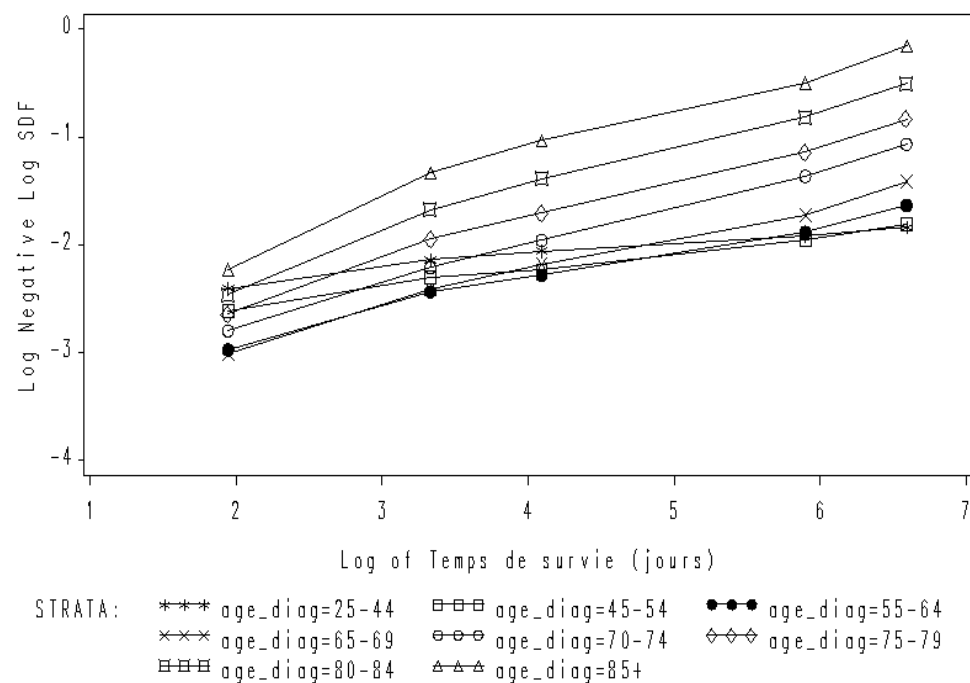
Année de l'hospitalisation (AN_DIAG) : L'année à laquelle l'AVC survient aurait un rôle important à jouer sur la survie. L'année de diagnostic (ici l'hospitalisation, car on ne connaît pas la date du diagnostic) a souvent été utilisée dans des études sur la survie à différents types de cancer (Dickman *et al.*, 2004; Verdecchia *et al.*, 2003, 2004).

Temps de suivi (FU) : Le temps de suivi est le temps depuis l'hospitalisation pour AVC. Cette variable a été créée par une stratification en cinq intervalles, qui ont déjà été définis. Il s'agit d'une variable importante dans l'étude des AVC, car les premiers jours suivant l'hospitalisation sont souvent critiques et les changements de risque y sont plus rapides.

Type d'AVC (TYPE_AVC) : La dernière covariable différencie trois types d'AVC : l'AVC de type ischémique (CIM-9 : 433, 434, 436), l'AVC de type hémorragique (CIM-9 : 430, 431, 432) et l'AVC de type autre (CIM-9 : 435, 437, 438). Il est important de distinguer les types d'AVC, car le comportement d'indicateurs, telle la survie, change pour chacun des types. Cela a été montré pour la survie observée lors d'une étude sur la mortalité après un AVC sévère en Amérique du Sud (Braga *et al.*, 2002). Au Québec, une étude sur l'évolution du taux d'incidence de la maladie a aussi fait la distinction entre les types d'AVC (Mayo *et al.*, 1991).

L'hypothèse de proportionnalité des risques dans le temps a d'abord été vérifiée pour chacune des covariables à l'aide d'une méthode graphique. Elle a semblé appropriée pour les covariables sexe, année de l'hospitalisation et type d'AVC. Seuls les risques des catégories 25-44, 45-54 et 55-64 de la covariable âge lors de l'hospitalisation ne sont pas proportionnels, comme le montre la figure 4.2. Mais selon Paul Dickman ([Dickman et al., 2004](#)), les excès de risque sont habituellement non proportionnels pour l'âge.

FIG. 4.2 – Vérification graphique de l'hypothèse de proportionnalité des risques pour la variable âge lors de l'hospitalisation.



Par ailleurs, seul le logiciel SAS, particulièrement les procédures GENMOD et NLP, a été employé pour la modélisation. Le programme utilisé, développé par [Dickman et Hakulinen \(2002\)](#), a été adapté aux AVC. Ce programme est présenté à l'annexe C. Le programme original est disponible en ligne : <http://www.pauldickman.com/teaching/index.php>.

Encore une fois, la méthode Ederer II a été utilisée pour estimer la probabilité de

survie attendue pour tous les modèles, car cette méthode donne de meilleures estimations de la survie attendue spécifique aux intervalles et cumulative pour les données sur les accidents vasculaires cérébraux (voir section 4.4.2).

En considérant les cinq covariables mentionnées ci-haut, voici les approches qui ont été utilisées pour estimer le modèle :

1. Modèle 1 : maximisation de la vraisemblance de l'équation (3.2) à l'aide de la procédure NLP. La vraisemblance à maximiser est donc

$$l(\beta) = \sum_{j=0}^J \sum_{k=1}^K \left\{ \delta_{jk} \ln[\lambda_k^*(j) + \exp(\beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc)] - y_{jk} \exp(\beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc) \right\}.$$

En SAS, nous devons seulement écrire cette vraisemblance et mentionner qu'il faut la maximiser (voir le programme en annexe).

2. Modèle 2 (équation (3.3)) : modèle linéaire généralisé avec erreur de structure poisson et offset $\ln(y_{jk})$ à l'aide de la procédure GENMOD. Ce modèle est caractérisé par

- $\delta_{jk} \sim \text{Poisson}(\mu_{jk} = y_{jk}[\lambda_k^*(j) + \exp(\beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc)])$;
- $Y_{jk} = \delta_{jk}$;
- $\mu_{jk} = E(Y_{jk}) = \delta_{jk}^* + y_{jk} \exp(\beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc)$;
- $g(\mu_{jk}) = \ln(\mu_{jk} - \delta_{jk}^*) = \ln(y_{jk}) + \beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc$.

3. Modèle 3 (équation (3.9)) : modèle linéaire généralisé avec erreur de structure poisson et offset $\ln[(t_j - t_{j-1})n_{jk}]$ à l'aide de la procédure GENMOD. Ce modèle est défini par :

- $\delta_{jk} \sim \text{Poisson}(\mu_{jk} = \delta_{jk}^* + (t_j - t_{j-1})n_{jk} \exp(\beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc))$;
- $g(\mu_{jk}) = \ln(\mu_{jk} - \delta_{jk}^*) = \ln((t_j - t_{j-1})n_{jk}) + \beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc$.

La différence entre le modèle 2 et le modèle 3 réside seulement dans l'offset, c'est-à-dire dans la mesure d'exposition. En effet, on utilise le temps à risque total pour tous les individus comme mesure d'exposition dans le modèle 2, alors que cette quantité est approximative dans le modèle 3.

4. Modèle 4 (équation (3.10)) : modèle linéaire généralisé avec erreur de structure binomiale à l'aide de la procédure GENMOD. Ce modèle est caractérisé par
- $\delta_{jk} \sim \text{Binomiale}(n_{jk}, 1 - p_{jk})$;
 - $Y_{jk} = \frac{ns_{jk}}{n_{jk}}$ où $ns_{jk} = 1 - \delta_{jk}$ est le nombre de sujets qui survivent à l'intervalle ;
 - $\mu_{jk} = E(Y_{jk}) = p_{jk}$;
 - $g(\mu_{jk}) = g(p_{jk}) = \ln \left[-\ln \frac{p_{jk}}{p_{jk}^*} \right] = \beta_0 + \beta_1 Fu + \beta_2 Sex + \beta_3 An_diag + \beta_4 Age_diag + \beta_5 Type_avc.$

En fait, ces quatre modèles sont très semblables, car ils sont des variantes du même modèle de base. Toutes ces approches supposent que l'excès de risque est constant à l'intérieur de chacun des intervalles de suivi et elles nécessitent toutes la pré-spécification des longueurs d'intervalles (Dickman *et al.*, 2004).

Les estimations obtenues pour ces quatre modèles sont présentées au tableau 4.6, sous forme d'excès de risque relatif par rapport aux catégories de référence de chacune des covariables, qui sont la période de suivi de 0 à 7 jours, les hommes, l'année d'hospitalisation 1990, la catégorie d'âge 25-44 ans et le type d'AVC ischémique.

Puisque trois des quatre modèles sont basés sur un modèle linéaire généralisé, il est possible d'utiliser la déviance comme mesure d'ajustement de ceux-ci. Sous l'hypothèse que le modèle fournit un ajustement adéquat des données, la déviance va suivre une distribution χ^2 avec degrés de liberté égaux au nombre de degrés de liberté résiduels du modèle (nombre d'observations - nombre de paramètres estimés). Ceci dit, le modèle 3 serait le meilleur des quatre modèles pour l'ajustement des données sur les accidents vasculaires cérébraux. Le modèle 1, quoique très précis, est rejeté face au modèle 3, car il est difficile à utiliser, mais aussi, parce qu'il ne permet pas d'estimer des termes d'interaction.

Quoiqu'il en soit, les résultats obtenus à l'aide du modèle 3 ne semblent pas corroborer ceux de la table 4.5 et semblent tout de même démontrer un mauvais ajustement (Déviance/dl = 2.4083). De possibles explications à ce mauvais ajustement sont l'omission d'importantes covariables et l'absence d'importants termes d'interactions. Dans ce cas-ci, l'explication la plus plausible serait celle de l'absence d'interactions. Une recherche du meilleur modèle (sur la base du modèle 3) incluant des interactions s'est

TAB. 4.6 – Estimations des ratios d'excès de risque ($\exp(\beta)$) pour quatre approches de modélisation de la survie relative pour les cas d'AVC survenus entre 1990 et 1992.

	Modèle 1		Modèle 2		Modèle 3		Modèle 4	
	RER	IC (95%)	RER	IC (95%)	RER	IC (95%)	RER	IC (95%)
Déviante			1962.7931		1693.0102		1864.1972	
D.l.			703		703		703	
Suivi 7-28	0.326	(0.307,0.346)	0.327	(0.308,0.347)	0.318	(0.299,0.338)	0.952	(0.895,1.012)
Suivi 28-60	0.119	(0.111,0.129)	0.120	(0.111,0.129)	0.111	(0.102,0.120)	0.494	(0.456,0.535)
Suivi 60-1 an	0.028	(0.027,0.030)	0.028	(0.027,0.030)	0.025	(0.023,0.027)	1.138	(1.062,1.218)
Suivi 1 an-2 ans	0.015	(0.014,0.016)	0.015	(0.014,0.016)	0.014	(0.013,0.016)	0.773	(0.707,0.844)
Femmes	0.994	(0.949,1.041)	0.995	(0.950,1.042)	1.026	(0.977,1.077)	1.021	(0.973,1.072)
Année d'hosp. 1991	0.932	(0.882,0.985)	0.933	(0.883,0.986)	0.932	(0.880,0.988)	0.929	(0.877,0.984)
Année d'hosp. 1992	0.896	(0.847,0.947)	0.898	(0.850,0.949)	0.896	(0.846,0.950)	0.893	(0.842,0.946)
Âge 45-54	1.296	(1.089,1.541)	1.299	(1.092,1.545)	1.284	(1.078,1.528)	1.304	(1.096,1.553)
Âge 55-64	1.588	(1.361,1.852)	1.593	(1.366,1.859)	1.542	(1.320,1.801)	1.582	(1.355,1.848)
Âge 65-69	1.957	(1.675,2.286)	1.961	(1.679,2.291)	1.857	(1.587,2.173)	1.924	(1.645,2.251)
Âge 70-74	2.750	(2.366,3.197)	2.756	(2.371,3.203)	2.538	(2.180,2.954)	2.675	(2.298,3.114)
Âge 75-79	3.385	(2.918,3.927)	3.390	(2.923,3.933)	3.060	(2.634,3.554)	3.267	(2.812,3.795)
Âge 80-84	4.588	(3.956,5.320)	4.599	(3.966,5.334)	3.996	(3.440,4.641)	4.355	(3.750,5.058)
Âge 85 et +	6.416	(5.530,7.445)	6.422	(5.535,7.452)	5.303	(4.561,6.165)	5.936	(5.106,6.901)
Type autre	0.338	(0.311,0.367)	0.337	(0.310,0.366)	0.312	(0.284,0.343)	0.305	(0.278,0.336)
Type hémorragique	2.680	(2.535,2.833)	2.682	(2.537,2.835)	2.581	(2.438,2.731)	2.702	(2.554,2.860)

donc effectuée. Cette recherche a été basée sur une méthode de sélection backward, débutant avec un modèle complexe et éliminant successivement les termes apportant la plus petite contribution au modèle, c'est-à-dire les moins significatifs. Le tableau 4.7 présente les étapes successives de la sélection des variables et interactions à inclure dans le modèle, la première étape étant le modèle avec toutes les interactions doubles. Le calcul du seuil observé du test est basé sur la différence de déviance entre le modèle proposé et le modèle précédent.

Comme nous pouvons le constater, tous les modèles proposés, lorsqu'ils sont comparés au modèle précédent, s'ajusteraient bien aux données.

Le modèle choisi est donc le modèle 9, c'est-à-dire :

TAB. 4.7 – Étapes pour l’ajustement du modèle de survie relative choisi.

Modèle	Déviante	d.l.	Déviante/d.l.	Différence	pvalue
(1) Toutes les interactions doubles	629.8190	612	1.0291		
(2) - fu*sex	630.9982	616	1.0243	1.1792 (dl=4)	0.88151
(3) - fu*sex - age_diag*an_diag	640.3082	630	1.0164	9.31 (dl=14)	0.81069
(4) - fu*sex - age_diag*an_diag - fu*an_diag	645.1089	638	1.0111	4.8007 (dl=8)	0.77865
(5) - fu*sex - age_diag*an_diag - fu*an_diag - an_diag*type_avc	649.0135	642	1.0109	3.9046 (dl=4)	0.41907
(6) - fu*sex - age_diag*an_diag - fu*an_diag - an_diag*type_avc - sex*an_diag	651.4302	644	1.0115	2.4167 (dl=2)	0.29869
(7) - fu*sex - age_diag*an_diag - fu*an_diag - an_diag*type_avc - sex*an_diag - sex*age_diag	663.5564	651	1.0193	12.1262 (dl=7)	0.09648
(8) - fu*sex - age_diag*an_diag - fu*an_diag - an_diag*type_avc - sex*an_diag - sex*age_diag - sex*type_avc	668.9773	653	1.0245	5.4209 (dl=2)	0.06651
(9) - fu*sex - age_diag*an_diag - fu*an_diag - an_diag*type_avc - sex*an_diag - sex*age_diag - sex*type_avc - sex	668.9803	654	1.0229	0.003 (dl=1)	0.95632

d = fu an_diag age_diag type_avc fu*age_diag fu*type_avc age_diag*type_avc

Il est étonnant de constater que la variable sexe n’apporte aucune contribution significative au modèle. Cet indicateur a aussi été rejeté du modèle lors d’une étude de modélisation de la survie au cancer colorectal ([Monnet et al., 1992](#)). Les estimations des paramètres ont tout de même été comparées dans le modèle avec la variable sexe et sans la variable sexe, afin de vérifier qu’il ne s’agissait pas d’un facteur confondant dans le modèle.

Les seuils des statistiques du chi-deux de l’analyse de Type3 pour le modèle choisi sont présentés au tableau 4.8. Il est à noter qu’il n’est pas justifié d’interpréter les statistiques LR des effets principaux fu, age_diag et type_avc, car ils figurent dans une interaction significative. Il existe donc des différences significatives dans la survie relative pour tous les termes du modèle présentés dans ce tableau. L’interaction entre l’âge lors de l’hospitalisation et le temps de suivi peut se traduire ainsi : l’excès de risque relatif pour l’âge diffère selon le temps de suivi. En d’autres mots, l’hypothèse de proportionnalité des excès de risque n’est pas appropriée.

La table 4.9 présente l’estimation des paramètres, ainsi que l’estimation des excès de risque relatif (ERR). Les ratios d’excès de risques relatifs ont été préférés aux risques relatifs comme mesure de comparaison des risques, car il s’agit de la mesure employée dans les études de survie au cancer de type populationnel ([Dickman et Hakulinen, 2002](#)). Les paramètres présentés sont ceux proposés par SAS. Il est toutefois un peu plus complexe d’obtenir les ERR. Pour un effet simple, l’excès de risque relatif est obtenu en prenant

TAB. 4.8 – Statistiques du rapport de vraisemblances pour l'analyse de Type 3.

Source	DF	Chi-Square	Pr > ChiSq
fu	4	824.89	<.0001
an_diag	2	13.36	0.0008
age_diag	7	594.95	<.0001
type_avc	2	372.76	<.0001
fu*age_diag	28	153.35	<.0001
fu*type_avc	8	616.39	<.0001
age_diag*type_avc	14	84.81	<.0001

l'exponentielle de l'estimation du paramètre β et son intervalle de confiance est donné par $[e^{\beta-1.96ET}, e^{\beta+1.96ET}]$, où ET est l'erreur-type du paramètre. Pour une interaction, l'ERR est égal à l'exponentielle d'un nouveau paramètre β^* défini par la somme des estimations des paramètres β des deux effets simples et de l'interaction. Par exemple, l'ERR de l'interaction entre le deuxième intervalle de suivi et le groupe d'âge 70 à 74 ans, est donné par l'exponentielle de la somme de l'estimation du paramètre de l'effet simple pour le deuxième intervalle, de l'estimation du paramètre de l'effet simple pour le groupe d'âge 70 à 74 ans et de l'estimation du paramètre de l'interaction entre ces deux effets simples. Plus explicitement, $ERR = \exp(-1.692 + 0.409 + 0.600) = 0.505$. L'intervalle de confiance, associé à l'ERR d'une interaction, peut être calculé par $[e^{\beta^*-1.96ET}, e^{\beta^*+1.96ET}]$ où $\beta^* = \ln(ERR)$ et ET est l'écart-type de ce paramètre β^* . Alors l'ET sera la racine carrée de la somme des variances des paramètres β , plus deux fois les covariances. La matrice des covariances peut être demandée en SAS avec l'option COVB de proc GENMOD. Donc pour l'interaction fu*age_diag avec fu=2 et age_diag=70-74, nous avons $Var(\beta^*) = 0.200^2 + 0.409^2 + 0.213^2 + 2(0.0121) + 2(-0.0386) + 2(-0.0154) = 0.1688$ et l'écart-type est 0.411. Tel qu'indiqué dans le tableau 4.9, l'intervalle de confiance de l'excès de risque relatif de l'interaction entre le deuxième intervalle de suivi et le groupe d'âge 70-74 ans est $[e^{-0.683-1.96(0.411)}, e^{-0.683+1.96(0.411)}] = [0.342, 0.745]$.

Les ERR des termes d'interactions s'interprètent comme l'excès de risque expérimenté par les sujets de cette catégorie, par rapport aux sujets faisant partie de la catégorie de référence jointe. Par exemple, après avoir ajusté pour les autres facteurs, les individus âgés entre 55 et 64 ans qui ont subi un AVC de type hémorragique ont expérimenté un excès de risque de mortalité de 623% de l'excès de mortalité expérimenté par les individus âgés de 25 à 44 ans, qui ont eux subi un AVC de type ischémique. De plus, cette différence est statistiquement significative au seuil de 5% (Test de Wald : $P[\chi_1^2 > (\text{paramètre}/\text{erreur standard})^2] = P[\chi_1^2 > (-1.830/0.200)^2] < 0.0001$). La significativité des résultats se reflète aussi dans les intervalles de confiance : si 1 n'est pas inclus dans

TAB. 4.9 – Estimations des paramètres et des ratios d'excès de risque ($\exp(\beta)$) pour le modèle final.

Paramètre	Niveau 1	Niveau 2	Paramètre SAS (ET)	ERR	IC (95%)
fu	2		-1.692 (0.200)	0.184	(0.124,0.272)
	3		-2.984 (0.307)	0.051	(0.028,0.092)
	4		-4.471 (0.258)	0.011	(0.007,0.019)
	5		-4.822 (0.329)	0.008	(0.004,0.015)
	1991		-0.065 (0.029)	0.937*	(0.885,0.992)
an_diag	1992		-0.106 (0.029)	0.900*	(0.850,0.953)
age_diag	45-54		-0.034 (0.220)	0.966	(0.628,1.488)
	55-64		-0.007 (0.198)	0.993	(0.673,1.463)
	65-69		0.118 (0.199)	1.125	(0.762,1.662)
	70-74		0.409 (0.194)	1.505	(1.029,2.202)
	75-79		0.637 (0.192)	1.891	(1.297,2.756)
	80-84		0.870 (0.192)	2.386	(1.637,3.477)
	85 et +		1.195 (0.192)	3.305	(2.269,4.184)
	type_avc	autre		-1.547 (0.360)	0.213
fu*age_diag	hémorragique		1.630 (0.187)	5.105	(3.539,7.364)
	2	45-54	0.139 (0.246)	0.204*	(0.127,0.328)
	2	55-64	0.601 (0.216)	0.333*	(0.224,0.496)
	2	65-69	0.664 (0.219)	0.403*	(0.270,0.599)
	2	70-74	0.600 (0.213)	0.505*	(0.342,0.745)
	2	75-79	0.814 (0.209)	0.785	(0.537,1.149)
	2	80-84	0.932 (0.209)	1.116	(0.764,1.628)
	2	85 et +	1.034 (0.209)	1.711*	(1.176,2.490)
	3	45-54	-0.257 (0.410)	0.038*	(0.019,0.073)
	3	55-64	0.530 (0.332)	0.085*	(0.054,0.134)
	3	65-69	0.949 (0.327)	0.147*	(0.096,0.225)
	3	70-74	0.998 (0.320)	0.207*	(0.138,0.310)
	3	75-79	1.062 (0.317)	0.277*	(0.186,0.412)
	3	80-84	1.289 (0.316)	0.438*	(0.296,0.648)
	3	85 et +	1.339 (0.317)	0.638*	(0.432,0.942)
	4	45-54	0.639 (0.298)	0.021*	(0.013,0.033)
	4	55-64	0.902 (0.273)	0.028*	(0.019,0.042)
	4	65-69	0.943 (0.277)	0.033*	(0.022,0.049)
	4	70-74	1.282 (0.268)	0.062*	(0.042,0.091)
	4	75-79	1.167 (0.267)	0.069*	(0.047,0.102)
	4	80-84	1.209 (0.268)	0.091*	(0.062,0.134)
	4	85 et +	0.931 (0.273)	0.096*	(0.065,0.142)
	5	45-54	0.440 (0.384)	0.012*	(0.007,0.021)
	5	55-64	0.871 (0.345)	0.019*	(0.013,0.029)
	5	65-69	1.104 (0.346)	0.027*	(0.018,0.041)
	5	70-74	1.020 (0.342)	0.034*	(0.022,0.050)
	5	75-79	0.981 (0.342)	0.041*	(0.027,0.061)
	5	80-84	0.988 (0.344)	0.052*	(0.034,0.078)
fu*type_avc	5	85 et +	0.927 (0.347)	0.067*	(0.044,0.102)
	2	autre	0.175 (0.150)	0.047*	(0.021,0.104)
	2	hémorragique	-0.643 (0.076)	0.494*	(0.285,0.857)
	3	autre	0.509 (0.168)	0.018*	(0.007,0.045)
	3	hémorragique	-0.841 (0.109)	0.111*	(0.054,0.229)
	4	autre	1.138 (0.134)	0.008*	(0.003,0.017)
	4	hémorragique	-1.386 (0.104)	0.015*	(0.007,0.029)
	5	autre	1.273 (0.145)	0.006*	(0.003,0.015)
	5	hémorragique	-2.120 (0.189)	0.005*	(0.002,0.011)
	age_diag*type_avc	45-54	autre	-0.249 (0.425)	0.160*
45-54		hémorragique	0.313 (0.223)	6.746*	(4.523,10.063)
55-64		autre	0.035 (0.370)	0.219*	(0.134,0.357)
55-64		hémorragique	0.207 (0.201)	6.234*	(4.212,9.226)
65-69		autre	-0.003 (0.369)	0.239*	(0.146,0.390)
65-69		hémorragique	0.174 (0.205)	6.835*	(4.573,10.214)
70-74		autre	-0.078 (0.362)	0.296*	(0.186,0.474)
70-74		hémorragique	0.078 (0.200)	8.306*	(5.591,12.341)
75-79		autre	-0.267 (0.363)	0.308*	(0.193,0.491)
75-79		hémorragique	-0.123 (0.198)	8.542*	(5.761,12.666)
80-84		autre	-0.252 (0.362)	0.395*	(0.247,0.630)
80-84		hémorragique	-0.326 (0.200)	8.793*	(5.907,13.090)
85 et +		autre	-0.196 (0.363)	0.578*	(0.362,0.924)
85 et +		hémorragique	-0.600 (0.206)	9.263*	(6.149,13.952)

*Résultat significatif au seuil de 5%.

l'intervalle, le résultat est significatif. Par ailleurs, les individus hospitalisés en 1992 ont expérimenté un excès de mortalité significativement inférieur de 10% à celui des individus diagnostiqués en 1990. En somme,

- les individus de 45 ans et plus qui ont survécu plus de 7 jours ont en général expérimenté un excès de risque inférieur aux individus de 25 à 44 ans qui sont décédés dans les 7 premiers jours suivant l'hospitalisation ;
- les individus qui ont subi un AVC de type autre ou de type hémorragique et qui ont survécu plus de 7 jours ont expérimenté un excès de risque inférieur à celui des individus qui ont subi un AVC de type ischémique et qui sont décédés dans les 7 jours suivant l'hospitalisation ;
- les individus de 45 ans et plus qui ont subi un AVC de type autre ont expérimenté un excès de risque inférieur à celui des individus de 25 à 44 ans qui ont subi un AVC de type ischémique, alors que les individus de la même tranche d'âge qui ont subi un AVC de type hémorragique ont expérimenté un excès de risque largement supérieur à celui des individus de moins de 45 ans qui ont subi un AVC de type ischémique ;
- les individus qui ont été hospitalisés en 1991 et 1992 ont expérimenté un excès de risque inférieur à celui des individus hospitalisés en 1990.

Il est toutefois difficile d'établir un parallèle entre les tableaux 4.5 et 4.9, c'est-à-dire entre l'estimation et la modélisation de la survie relative. En effet, la survie relative est multiplicative, alors que le modèle de risques utilisé est additif. En fait, le moyen le plus simple pouvant être employé ici est de calculer un excès de risque relatif à partir des résultats marginaux du tableau 4.5. On constate d'ailleurs à l'aide de cette méthode que les résultats marginaux ne seraient pas un bon résumé des résultats de la modélisation. Par exemple, pour l'interaction $fu \cdot \text{age_diag}$, où $fu=2$ et $\text{age_diag}=85+$, l'excès de risque relatif calculé après la modélisation (tableau 4.9) est de 1.711. Si cet excès de risque est calculé à partir des résultats marginaux (tableau 4.5), on aura

$$\begin{aligned} ERR &= \frac{1 - \text{survie relative des } 85+ \text{ à } 28 \text{ jours de suivi}}{1 - \text{survie relative des } 25-44 \text{ ans à } 7 \text{ jours de suivi}} \\ &= \frac{1 - 0.862}{1 - 0.914} = 1.605, \end{aligned}$$

où le numérateur correspond au risque de décès chez les plus de 85 ans pour le deuxième intervalle de suivi, et le dénominateur correspond au risque de décès pour la catégorie de référence, c'est-à-dire chez les 25 à 44 ans pour le premier intervalle de suivi. On voit que l'excès de risque relatif de cette interaction, calculé à partir des résultats marginaux, diffère légèrement de celui calculé à partir des résultats de la modélisation, mais qu'il est du même ordre de grandeur. Cependant, plus la catégorie d'âge et le temps de suivi augmentent, plus l'écart entre les deux ERR se prononce. Chez les 85 ans et plus au

cinquième intervalle de suivi, l'ERR calculé à partir des résultats de la modélisation est de 0.067, alors qu'il est de 1.279 lorsqu'il est calculé à partir des résultats marginaux. Cette différence peut probablement s'expliquer par la non sommabilité des résultats marginaux. Il est donc préférable d'estimer la survie relative par modélisation lorsque la présence d'interactions significatives est soupçonnée.

Enfin, les postulats de normalité et d'homoscédasticité des résidus ont été vérifiés pour le modèle final. Pour une régression Poisson, les résidus suivent une loi normale sous H_0 . Cependant, les tests formels de normalité, avec un seuil observé inférieur à 5%, ont indiqué que l'hypothèse de normalité des résidus est rejetée. En ce qui concerne la relation entre les résidus du modèle et les valeurs prédites, un léger patron a pu être décelé, suggérant ainsi que l'hypothèse d'homoscédasticité n'est pas non plus respectée.

4.6 Discussion

Ce mémoire sur différentes approches de modélisation de la survie relative a permis de comparer les caractéristiques mathématiques de chacune, ainsi que leurs résultats respectifs sur des données sur les accidents vasculaires cérébraux. Le but était de comparer ces méthodes et de déterminer laquelle serait la plus appropriée pour les AVC.

Chacune des approches présentées pour l'estimation du modèle de survie relative a produit des estimations similaires. Cela n'a rien de surprenant, car elles estiment le même modèle fondamental à l'aide de la même méthode de maximisation (maximum de vraisemblance).

Le modèle proposé par Estève et *al.* était, à prime abord, supérieur du point de vue théorique, car tous les individus peuvent être inclus dans le modèle sur la base de données individuelles. En effet, il est préférable d'utiliser des méthodes basées sur l'estimation individuelle, surtout lorsque l'échantillon est de petite taille ou que le nombre d'individus dans chacune des strates étudiées est petit. Ces méthodes sont plus puissantes que celles qui sont basées sur des données groupées. Cependant, l'estimation des paramètres ne s'est pas avérée très différente des autres méthodes. De plus, cette approche est plus difficile et beaucoup moins rapide d'application que les autres. Compte-tenu de cela, il est préférable d'estimer le modèle dans le cadre des modèles linéaires généralisés. Cette façon de faire bénéficie de tous les avantages de la méthode d'Estève

et *al.* et apporte l'avantage additionnel de pouvoir travailler à l'aide d'une régression et ainsi pouvoir compter sur des tests d'ajustement et la possibilité d'inclure des termes d'interactions (Dickman *et al.*, 2004).

Dans la méthode d'Hakulinen et de Tenkanen, les individus sont groupés en strates et par conséquent, la proportion de survie attendue pour la strate est calculée comme la moyenne des proportions individuelles des sujets de la strate. Un inconvénient de cette méthode est qu'elle groupe les sujets avec survie attendue hétérogène et elle assume que tous les individus ont la même probabilité de survie jusqu'à la fin de l'intervalle de temps (Monnet *et al.*, 1992). Cette approximation peut être satisfaisante quand les intervalles de temps choisis sont courts. Un autre désavantage de l'utilisation de strates est la perte de puissance, ce qui peut être un obstacle dans les études avec un petit nombre d'individus ou pour la survie à long terme.

Le modèle qui a été choisi ici pour l'application aux AVC est un modèle linéaire généralisé avec une erreur de structure Poisson et avec une approximation en termes de taux du nombre de décès attendu. Ce modèle était celui parmi les quatre modèles comparés qui s'ajustait le mieux aux données. Des termes d'interaction ont ensuite été ajoutés au modèle pour améliorer son ajustement et ce, à l'aide d'une méthode de sélection backward. Cette méthode a été préférée aux méthodes de sélection forward et stepwise, car elle était plus facile d'application. En fait, la sélection de modèle, qui est disponible en SAS pour plusieurs procédures (REG, PHREG, LOGISTIC, etc.), n'est pas encore disponible avec GENMOD et ce, contrairement au logiciel GLIM qui permet la sélection de modèle pour une régression Poisson. La méthode a donc dû être appliquée manuellement.

Les résultats obtenus pour la modélisation ne peuvent cependant pas être comparés aux résultats d'études antérieures, car ils n'ont pas été standardisés pour l'âge. De toute façon, le but de ce mémoire n'était pas de faire une comparaison interprovinciale ou internationale, mais bien une comparaison de certaines méthodes de modélisation.

Certains auraient pu se demander s'il était réellement utile de tenir compte de l'âge dans le modèle de régression de survie relative, puisqu'il est déjà pris en compte dans les tables de mortalité. L'âge est effectivement pris en compte dans les tables de mortalité, mais seulement en ce qui concerne la mortalité par autres causes directement liées au vieillissement. En fait, l'âge peut encore intervenir dans l'excès de risque. Il est donc

justifié qu'il soit inclus dans le modèle.

Par ailleurs, la base de données sur les AVC ne contenait pas de données censurées, car on supposait que les individus non décédés étaient toujours vivants. Toutes les méthodes présentées peuvent cependant être adaptées pour tenir compte de la censure à droite. Par exemple, à la section 3.3, n_{jk} , le nombre d'individus en vie au début de l'intervalle j de la strate k , pourrait être remplacé par $l_{jk} = n_{jk} - w_{jk}/2$, le nombre effectif de sujets à risque pendant l'intervalle j de la strate k , où w_{jk} est le nombre d'individus censurés pendant l'intervalle j de la strate k .

En somme, la modélisation de la survie relative est d'une très grande utilité dans les études analytiques. Elle permet d'abord de prendre en compte plusieurs covariables et interactions, mais elle permet aussi d'éliminer une partie de la variabilité aléatoire, moyennant quelques hypothèses sur le lien entre les facteurs et le processus de décès ([Hédelin, 2000](#)).

Conclusion

Dans ce mémoire, différentes approches d'estimation de la survie relative ont été comparées. L'estimation du rapport de survie relative s'est avérée être un bon moyen pour décrire la survie lorsque le nombre de covariables est assez restreint. Dans le cas contraire, la modélisation à l'aide d'un modèle de risques additifs est une bonne alternative pour tenir compte de plusieurs variables explicatives.

L'étude de type populationnel sur les accidents vasculaires cérébraux a permis l'application des notions théoriques abordées dans ce mémoire. Il a entre autre été montré qu'une approche basée sur une régression Poisson est supérieure dans le cas des AVC. Par ailleurs, il a été étonnant de constater que le sexe ne faisait pas partie du modèle final choisi. Il n'existerait donc pas de différence significative entre l'excès de risque chez les hommes et chez les femmes, après correction pour d'autres facteurs, pour les années 1990 à 1992. Il en aurait peut-être été autrement si l'échantillon d'années avait été plus grand.

Par ailleurs, il aurait été intéressant d'intégrer d'autres covariables dans le modèle, comme par exemple, le niveau socio-économique, le niveau socioprofessionnel, la région, etc. D'ailleurs, plusieurs auteurs ([Dickman *et al.*, 1998, 1997](#); [Coleman *et al.*, 1999](#)) sont unanimes sur la nécessité d'utiliser des tables prenant en compte la région et le statut socio-économique. Cela aurait, par contre, exigé l'utilisation de tables de mortalité adaptées, qui auraient dû être estimées pour l'ensemble du Québec. En ce qui concerne la stratification régionale, une table de mortalité par région aurait pu être estimée en corrigeant les tables de l'ensemble de la province.

De plus, d'autres modèles auraient pu être proposés et comparés pour estimer la survie relative aux accidents vasculaires cérébraux, notamment des méthodes bayésiennes, des modèles paramétriques, des modèles basés sur des fonctions B-splines, etc. Certains

de ces modèles ont déjà été abordés et comparés par [Giorgi \(2002\)](#). Mais la plupart de ces méthodes sont encore en développements et nécessitent l'utilisation de logiciels particuliers. Il pourrait alors s'agir d'une avenue à explorer dans le choix d'une approche d'estimation de la survie relative aux accidents vasculaires cérébraux pour des études futures.

Bibliographie

- AIDE EN LIGNE DE SAS (1999). *Aide en ligne de SAS*. SAS Institute Inc., Cary, NC, 8ième édition.
- ANDERSEN, P., BORCH-JOHNSEN, K., DECKER, T., GREEN, A., HOUGAARD, P., KEIDING, N. et KREINER, S. (1985). A cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics*, 41(4):921–932.
- ANDERSEN, P. et VAETH, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics*, 45(2):523–535.
- BERKSON, J. (1942). The calculation of survival rates. In WALTERS, W., GRAY, H. K. et T., P. J., éditeurs : *Carcinoma and other malignant lesions of the stomach*, pages 467–484.
- BERKSON, J. et GAGE, R. (1950). Calculation of survival rates for cancer. *Proceeding of the Staff Meeting of the Mayo Clinic*, 25:270–286.
- BOUCHARD, C. et LOUCHINI, R. (1999). Surveillance de la mortalité au québec : 1976-1997. Rapport technique, Direction de la santé publique du Québec, ministère de la Santé et des Services sociaux du Québec, Québec.
- BOUCHARD, C. et LOUCHINI, R. (2001). Surveillance de la mortalité au québec : 1977-1998. Rapport technique, Direction de la santé publique du Québec, ministère de la Santé et des Services sociaux du Québec, Québec.
- BRAGA, P., IBARRA, A., REGA, I., KETZOIAN, C., PEBET, M., SERVENTE, L. et BENZANO, D. (2002). Prediction of early mortality after acute stroke. *Journal of Stroke and Cerebrovascular Diseases*, 11(1):15–22.
- BRAY, D., BRANCKER, A. et ADAMS, O. (1990). Tables de mortalité, Canada et provinces, 1985-1987. Rapport technique 2(4) Suppl.13, Statistique Canada. Rapport sur la santé.
- BRESLOW, N. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43:45–58.

- BRESLOW, N. et DAY, N. (1987). *Statistical Methods in Cancer Research : Volume II - The Design and Analysis of Cohort Studies*. IARC Scientific Publications No. 82. International Agency for Research on Cancer, Lyon.
- BRESLOW, N., LUBIN, J., MAREK, P. et LANGHOLZ, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78:1–12.
- BUCKLEY, J. (1984). Additive and multiplicative models for relative survival rates. *Biometrics*, 40(1):51–62.
- CHEUVART, B. et RYAN, L. (1991). Adjusting for age-related competing mortality in long-term cancer clinical trials. *Statistics in Medicine*, 10:65–77.
- CHIANG, C. (1968). *Introduction to Stochastic Processes in Biostatistics*. John Wiley and Sons, New York.
- COLEMAN, M., BABB, P., DAMIECKI, P., GROSCLAUDE, P., HONJO, S., JONES, J., KNERER, G., PITARD, A., QUINN, M., SLOGGET, A. et DE STAVOLA, B. (1999). Cancer survival trends in England and Wales, 1971-1995 : deprivations and NHS region. *In Studies in medical and population subjects #61*. Office for National Statistics.
- COX, D. (1972). Regression models and life table (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–202.
- CROWDER, J. (2001). *Classical competing risks*. Chapman and Hall, London.
- DICKMAN, P., AUVINEN, A., VOUTILAINEN, E. et HAKULINEN, T. (1998). Measuring social class differences in cancer patient survival : Is it necessary to control for social class differences in general population mortality ? A Finnish population-based study. *Journal of Epidemiology and Community Health*, 52:727–734.
- DICKMAN, P., GIBBERD, R. et HAKULINEN, T. (1997). Estimating potential savings in cancer deaths by eliminating regional social class variation in cancer survival in the Nordic countries. *Journal of Epidemiology and Community Health*, 51:289–298.
- DICKMAN, P. et HAKULINEN, T. (2002). Population-based cancer survival analysis. Notes de cours, Toronto.
- DICKMAN, P., SLOGGETT, A., HILLS, M. et HAKULINEN, T. (2004). Regression models for relative survival. *Statistics in Medicine*, 23:51–64.
- EDERER, F., AXTELL, L. et CUTLER, S. (1961). The relative survival rate : A statistical methodology. *National Cancer Institute Monograph*, 6:101–121.

- EDERER, F. et HEISE, H. (1959). The effect of eliminating deaths from cancer in general population survival rates. Rapport technique, National Cancer Institute. methodological note 11, End result Evaluation Section.
- ELLISON, L., GIBBONS, L. et LE GROUPE D'ANALYSE DE LA SURVIE AU CANCER AU CANADA (2001). Taux relatifs de survie à cinq ans - cancers de la prostate, du sein, du côlon et du rectum, et du poumon. *Rapports sur la santé*, 13(1):1–12.
- ESTÈVE, J., BENHAMOU, E., CROASDALE, M. et RAYMOND, L. (1990). Relative survival and the estimation of net survival : elements for further discussion. *Statistics in medicine*, 9:529–538.
- FONDATION DES MALADIES DU COEUR DU CANADA (1999). Le nouveau visage des maladies cardiovasculaires et des accidents vasculaires cérébraux au Canada. Rapport technique, Fondation des maladies du coeur du Canada, Ottawa, Canada.
- FOUCHER, P., COUDERT, B., DRAMAIS-MARCEL, D., ARVEUX, P., CAMUS, P. et JEANNIN, L. (1994). Les apports de la survie relative par rapport à la survie classique. À propos du cancer bronchique primitif. *Bulletin of Cancer*, 81:857–865.
- GIORGI, R. (2002). *Analyses comparatives des méthodes de survie et extensions d'un modèle régressif de survie relative : prise en compte de la non proportionnalité des risques par des fonctions B-splines et développement d'une méthode d'analyse bayésienne*. Thèse de doctorat, Université d'Aix-Marseille, Marseille, France.
- HAKULINEN, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38(4):933–942.
- HAKULINEN, T. et TENKANEN, L. (1987). Regression analysis of relative survival rates. *Applied Statistics*, 36:309–317.
- HÉDELIN, G. (2000). *Les modèles de survie relative et leurs applications*. Thèse de doctorat, Université de Strasbourg, Strasbourg, France.
- KAPLAN, E. et MEIER, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American statistical association*, 53:457–481.
- KLEIN, J. et MOESCHBERGER, M. (1997). *Survival Analysis : Techniques for Censored and Truncated Data*. Springer, New York.
- LOUCHINI, R. (2002). La survie au cancer pour les nouveaux cas déclarés au Québec en 1992. Rapport technique, Institut National de Santé Publique du Québec, Québec.
- LOUCHINI, R. et BEAUPRÉ, M. (2003). La survie reliée au cancer pour les nouveaux cas déclarés au Québec, de 1984 à 1998 - Survie observée et survie relative. Rapport technique, Institut National de Santé Publique du Québec, Québec.

- MARUBINI, E., MEZZANOTTE, G., BORACCHI, P. et VERONESI, U. (1990). Long term survival analysis : breast cancer and age at diagnostic. *Statistics in Medicine*, 9:737–748.
- MAYO, N., GOLDBERG, M., LEVY, A., DANYS, I. et KORNER-BITENSKY, N. (1991). Changing rates of stroke in the province of Quebec, Canada : 1981-1988. *Stroke*, 22:590–595.
- MERCK & CO. (2004). *The Merck Manual of Diagnosis and Therapy*, Seventeenth édition. Disponible en ligne :
<http://www.merck.com/mrkshared/mmanual/section14/chapter174/174a.jsp>.
- MILLAR, W. et DAVID, P. (1995). Tables de mortalité, Canada et provinces, 1990-1992. Rapport technique 84-537 au catalogue, Statistique Canada, Ottawa.
- MONNET, E., BOUTRON, M., ARVEUX, P., MILAN, C. et FAIVRE, J. (1992). Different multiple regression models for estimating survival : use in a population-based series of colorectal cancers. *Journal of Clinical Epidemiology*, 45:267–273.
- POCOCK, S., GORE, S. et KERR, G. (1982). Long term survival analysis : the curability of breast cancer. *Statistics in medicine*, 1:93–104.
- SASIENI, R. (1996). Proportional excess hazard. *Biometrika*, 83:127–141.
- SERVICE VIE (2003). *La clé des maux : Accident vasculaire cérébral (AVC)*. Disponible en ligne :
http://www.servicevie.com/02Sante/Cle_des_maux/A/maux24.html.
- THERNEAU, T. et OFFORD, J. (1999). Expected survival based on hazard rates (updated). Rapport technique 63, Mayo Clinic.
- VERDECCHIA, A., CORAZZIARI, I., GATTA, G., LISI, D., FAIVRE, J., FORMAN, D. et the EURO CARE WORKING GROUP (2004). Explaining gastric cancer survival differences among european countries. *International Journal of Cancer*, 109:737–741.
- VERDECCHIA, A., MARIOTTO, A., GATTA, G., BUSTAMANTE-TEIXEIRA, M. et AJIKI, W. (2003). Comparison of stomach cancer incidence and survival in four continents. *European Journal of Cancer*, 39:1603–1609.
- VOUTILAINEN, E., DICKMAN, P. et HAKULINEN, T. (2000). SURV2 :Relative Survival Analysis Program. Finnish Cancer Registry : Helsinki. Disponible en ligne :
<http://www.cancerregistry.fi/surv2/>.
- ZAHL, P. (1993). Regression modelling for long-term survival of cancer, part II : a linear non-parametric regression model. Statistical research report 93-2, University of Oslo, Norvège.

ZAHL, P. (1995). A proportional regression model for 20 year survival of colon cancer in Norway. *Statistics in Medicine*, 14:1249–1261.

Annexe A

Tables québécoises de mortalité de 1986 et 1991

Q_H : Probabilité de décès chez les hommes

Q_F : Probabilité de décès chez les femmes

Sources : Tables de mortalité, Canada et provinces 1985-1987 et 1990-1992 ([Bray et al., 1990](#); [Millar et David, 1995](#))

ÂGE	1986		1991	
	Q_H	Q_F	Q_H	Q_F
0	0,00802	0,00622	0,00657	0,00516
1	0,00062	0,00058	0,00047	0,00041
2	0,00051	0,00042	0,00041	0,00030
3	0,00047	0,00032	0,00029	0,00026
4	0,00034	0,00029	0,00022	0,00019
5	0,00027	0,00026	0,00019	0,00016
6	0,00024	0,00023	0,00018	0,00015
7	0,00023	0,00020	0,00017	0,00014
8	0,00021	0,00018	0,00016	0,00013
9	0,00020	0,00016	0,00016	0,00013
10	0,00022	0,00016	0,00019	0,00013
11	0,00025	0,00016	0,00019	0,00012
12	0,00031	0,00017	0,00028	0,00015
13	0,00043	0,00020	0,00038	0,00017
14	0,00059	0,00024	0,00051	0,00020
15	0,00077	0,00028	0,00066	0,00023

ÂGE	1986		1991	
	Q_H	Q_F	Q_H	Q_F
16	0,00095	0,00033	0,00080	0,00026
17	0,00109	0,00036	0,00092	0,00029
18	0,00119	0,00037	0,00100	0,00031
19	0,00128	0,00037	0,00107	0,00033
20	0,00136	0,00037	0,00113	0,00035
21	0,00141	0,00037	0,00118	0,00037
22	0,00144	0,00037	0,00121	0,00038
23	0,00144	0,00038	0,00124	0,00039
24	0,00140	0,00039	0,00125	0,00040
25	0,00136	0,00041	0,00125	0,00040
26	0,00132	0,00043	0,00125	0,00041
27	0,00130	0,00045	0,00126	0,00042
28	0,00131	0,00048	0,00128	0,00044
29	0,00135	0,00051	0,00131	0,00046
30	0,00139	0,00054	0,00134	0,00048
31	0,00144	0,00058	0,00137	0,00051
32	0,00148	0,00062	0,00141	0,00055
33	0,00152	0,00065	0,00145	0,00059
34	0,00156	0,00069	0,00150	0,00064
35	0,00161	0,00072	0,00156	0,00069
36	0,00167	0,00077	0,00162	0,00075
37	0,00175	0,00083	0,00170	0,00082
38	0,00186	0,00091	0,00177	0,00090
39	0,00198	0,00099	0,00184	0,00098
40	0,00212	0,00109	0,00192	0,00107
41	0,00229	0,00121	0,00204	0,00118
42	0,00251	0,00135	0,00221	0,00130
43	0,00277	0,00152	0,00242	0,00143
44	0,00305	0,00171	0,00268	0,00157
45	0,00338	0,00191	0,00298	0,00173
46	0,00375	0,00214	0,00331	0,00191
47	0,00419	0,00238	0,00367	0,00209
48	0,00468	0,00264	0,00405	0,00228
49	0,00521	0,00291	0,00446	0,00248
50	0,00581	0,00321	0,00491	0,00270
51	0,00648	0,00352	0,00542	0,00294
52	0,00725	0,00385	0,00602	0,00322
53	0,00813	0,00420	0,00668	0,00353

ÂGE	1986		1991	
	Q_H	Q_F	Q_H	Q_F
54	0,00908	0,00455	0,00740	0,00385
55	0,01013	0,00493	0,00820	0,00421
56	0,01126	0,00536	0,00911	0,00462
57	0,01249	0,00586	0,01016	0,00510
58	0,01378	0,00641	0,01136	0,00565
59	0,01513	0,00701	0,01268	0,00628
60	0,01659	0,00767	0,01412	0,00695
61	0,01821	0,00841	0,01566	0,00768
62	0,02003	0,00924	0,01729	0,00844
63	0,02203	0,01015	0,01894	0,00920
64	0,02417	0,01112	0,02062	0,00996
65	0,02650	0,01219	0,02243	0,01078
66	0,02907	0,01339	0,02447	0,01173
67	0,03191	0,01476	0,02684	0,01286
68	0,03504	0,01622	0,02953	0,01412
69	0,03843	0,01777	0,03248	0,01547
70	0,04205	0,01949	0,03570	0,01700
71	0,04590	0,02147	0,03920	0,01876
72	0,04996	0,02382	0,04300	0,02085
73	0,05408	0,02645	0,04705	0,02323
74	0,05829	0,02932	0,05135	0,02584
75	0,06278	0,03252	0,05596	0,02874
76	0,06775	0,03613	0,06096	0,03199
77	0,07341	0,04026	0,06642	0,03565
78	0,07975	0,04480	0,07227	0,03961
79	0,08664	0,04970	0,07845	0,04383
80	0,09408	0,05509	0,08509	0,04847
81	0,10207	0,06112	0,09227	0,05371
82	0,11061	0,06794	0,10011	0,05971
83	0,11969	0,07544	0,10853	0,06636
84	0,12933	0,08354	0,11747	0,07355
85	0,13952	0,09236	0,12702	0,08144
86	0,145088	0,095873	0,13353	0,08574
87	0,156990	0,106302	0,14464	0,09499
88	0,169708	0,117706	0,15641	0,10507
89	0,183179	0,129976	0,16881	0,11588
90	0,197493	0,143282	0,18189	0,12758
91	0,212773	0,157792	0,19571	0,14025

$\hat{\text{ÂGE}}$	1986		1991	
	Q_H	Q_F	Q_H	Q_F
92	0,229119	0,173667	0,21032	0,15407
93	0,237319	0,181456	0,22568	0,16892
94	0,237308	0,181038	0,24176	0,18475
95	0,242906	0,186609	0,25862	0,20166
96	0,267922	0,212378	0,27630	0,21979
97	0,326082	0,272574	0,29485	0,23929
98	0,437693	0,388820	0,31424	0,26006
99	0,592666	0,551837	0,33445	0,28203
100	0,758952	0,729521	0,35551	0,30532
101	.	.	0,37748	0,33007
102	.	.	0,40041	0,35644
103	.	.	0,42428	0,38432
104	.	.	0,44904	0,41366
105	.	.	0,47475	0,44457

Annexe B

Liste des variables provenant du fichier sur les accidents vasculaires cérébraux

- annee* : année de l'hospitalisation
- clsc : CLSC de l'établissement
- codest : code de destination
- coprov : code de provenance
- dc48h : décès dans les 48 heures suivant l'hospitalisation
- dctype* : type de décès
- diagp : diagnostic principal
- diags1-15 : diagnostics secondaires 1 à 15
- dtadm* : date d'hospitalisation
- dtddc* : date de décès provenant du fichier des décès
- dtder* : date de décès provenant du fichier de la RAMQ
- dtnais* : date de naissance provenant du fichier MED-ECHO
- dtnaisdc* : date de naissance provenant du fichier des décès
- dtnaisr* : date de naissance provenant du fichier de la RAMQ
- dtsort* : date de sortie de l'hôpital
- namb* : numéro identifiant personnel
- sexe* : sexe provenant du fichier MED-ECHO
- sexedc* : sexe provenant du fichier de décès
- sexer* : sexe provenant du fichier de la RAMQ
- tydest : type de destination
- tyetab : type d'établissement
- typrov : type de provenance

* variables utilisées lors des analyses

Annexe C

Programme SAS pour la modélisation de la survie relative

A) Création des données de survie

```
/******  
SURVIVAL.SAS  
  
Estime la survie relative et produit des fichiers de donnees  
output qui peuvent etre utilises pour modeliser les modeles de  
survie relative.  
  
Eve-Marie Castonguay Mai 2004  
*****/  
title; footnote;  
  
title1 "Cas d'AVC diagnostiques au Quebec entre 1990 et 1992";  
  
libname survie 'C:\Documents and Settings\emcastonguay\Mes  
documents\AVC'; options fmtsearch=(survival)  
orientation=landscape;  
  
%let popmort=survie.popmort ;  
%let patdata=survie.avc ;  
%let individ=survie.individ ;  
%let grouped=survie.grouped ;  
  
%let vars = sex an_diag age_diag type_avc;  
%let lastvar = age_diag ;  
%let formats = sex sex. age_diag age_diag.;
```

```

data survie.individ1;
  set &patdata;
  id+1;
  survtime = survtime + 0.5;
  drop dctype sexe diagp dtnaidsdc sexedc dtdcdc sexer dtnaistr
      dtdcr diagnostic sexe1;
run;

%include 'C:\Documents and Settings\emcastonguay\Mes documents\AVC\
toronto2002\split.sas';

%split (
data=survee.individ1, out=survee.individ2, origin = 0, exit =
survtime, event = d, scale=1/365.25,
cuts = %str( 0,7/365.25,28/365.25,60/365.25,365.25/365.25,730.5/365.25 ) ) ;

data survie.individ3;
  set survie.individ2;

  _age=floor(((dtadm-datenais)+left)/365.25); *Si intervalles en jours;
  _year=put(floor(dtadm+left),year2.);

  if 83 < _year < 89 then _period=1986;
    else if 88 < _year < 94 then _period=1991;
    else _period=.;

  range=put(left,4.1) || ' - ' || left(put(right,4.1));
  sex2=sex-1;
  fu2=(fu=2);
  fu3=(fu=3);
  fu4=(fu=4);
  fu5=(fu=5);

  drop entry left right;
run;

proc sort data=survee.individ3;
  by sex _period _age;
run;

proc sort data=&popmort;
  by sex _period _age;
run;

data &individ (drop=an_diag rename=(an_diag_numeric=an_diag));
/*merger la table de mortalite*/
  merge survie.individ3(in=a) &popmort(in=b rename=(prob=temp));
  by sex _period _age;

```

```

    if a;
    prob=temp**length;
    e=-log(prob)*risk;
    an_diag_numeric=an_diag+0;

    drop temp;
    label
    e='Nombre attendu de deces'
    d="Indicateur pour deces pendant l'intervalle"
    w="Indicateur pour censure pendant l'intervalle"
    risk="Temps (annees) a risque pendant l'intervalle"
    length="Longueur de l'intervalle (potentielle et non actuelle)"
    log_risk='ln(risque)'
    prob='Probabilite de survie attendue'
    _age='Age atteint'
    _year='Annee de calendrier (mise à jour)'
    _period="Categorie d'annee 5-annees (meme que popmort)"
    fu='Intervalle de suivi'
    sex='Sexe';
run;

proc sort data=&individ;
    by id;
run;

proc summary data=&individ nway; /* Agreger les donnees pour
produire la table de survie */
    var d w prob risk e;
    id range length;
    class &vars fu; /* fu doit etre la derniere variable de cette liste */
    output out=survie.grouped1(drop=_type_ rename=(_freq_=n))
        sum(d w risk e)=d w y_exact e_exact mean(prob)=p_star;
    format &formats ;
run;

data &grouped ;
    retain cp cp_star cr 1;
    set survie.grouped1;

    if fu=1 then do;
        cp=1; cp_star=1; cr=1; se_temp=0;
    end;
    n_prime=n-w/2;
    ns=n_prime-d;
    p=1-d/n_prime;
    r=p/p_star;
    cp=cp*p;
    cp_star=cp_star*p_star;
    cr=cp/cp_star;

```

```

log_risk=log(n_prime-d/2);
log_personnes_temps=log(length*n);
ln_exact=log(y_exact);
e=n_prime*(1-p_star);
excess=(d-e)/y_exact;
se_p=sqrt(p*(1-p)/n_prime);
se_r=se_p/r;
/*Composante de la SE de la survie cumulative*/
se_temp+d/(n_prime*(n_prime-d));
se_cp=cp*sqrt(se_temp);
se_cr=se_cp/cp_star;

/* Calcul de l'IC sur l'échelle log-risque et transformation inverse */
/* D'abord pour les estimations d'intervalles-spezifiques */
if se_p ne 0 then do;
  /* SE sur l'échelle log-risque utilisant
     une approximation en serie de Taylor */
  se_lh_p=sqrt( se_p**2/(p*log(p))**2 );
  /* Limites de confiance sur l'échelle log-risque */
  lo_lh_p=log(-log(p))+1.96*se_lh_p;
  hi_lh_p=log(-log(p))-1.96*se_lh_p;
  /* Limites de confiance sur l'échelle de survie (survie observee) */
  lo_p=exp(-exp(lo_lh_p));
  hi_p=exp(-exp(hi_lh_p));
  /* Limites de confiance pour le taux de survie cumulative
     correspondant */
  lo_r=lo_p/p_star;
  hi_r=hi_p/p_star;

  drop se_lh_p lo_lh_p hi_lh_p;
  format lo_p hi_p lo_r hi_r 8.5;
  label
  lo_p='Lower 95% CI for P'
  hi_p='Upper 95% CI for P'
  lo_r='Lower 95% CI for R'
  hi_r='Upper 95% CI for R'
  ;
end;

/* Maintenant pour les estimations cumulatives */
if se_cp ne 0 then do;
  /* SE sur l'échelle log-risque utilisant une approximation
     en serie de Taylor */
  se_lh_cp=sqrt( se_cp**2/(cp*log(cp))**2 );
  /* Limites de confiance sur l'échelle log-risque */
  lo_lh_cp=log(-log(cp))+1.96*se_lh_cp;
  hi_lh_cp=log(-log(cp))-1.96*se_lh_cp;
  /* Limites de confiance sur l'échelle de survie (survie observee) */
  lo_cp=exp(-exp(lo_lh_cp));

```

```

hi_cp=exp(-exp(hi_lh_cp));
/* Limites de confiance pour les taux de
survie relative correspondants */
lo_cr=lo_cp/cp_star;
hi_cr=hi_cp/cp_star;

drop se_lh_cp lo_lh_cp hi_lh_cp;
format lo_cp hi_cp lo_cr hi_cr 8.5;
label
lo_cp='Lower 95% CI for CP'
hi_cp='Upper 95% CI for CP'
lo_cr='Lower 95% CI for CR'
hi_cr='Upper 95% CI for CR'
;
end;

drop se_temp;
label
range='Intervalle'
fu='Intervalle'
n='En vie au debut'
n_prime='Nombre effectif a risque'
ns="Nombre survivant a l'intervalle"
e='Nombre attendu de deces'
d='Deces'
w='Perdus de vue'
p='Survie observee pour un intervalle specifique'
cp='Survie observee cumulative'
r='Survie relative pour un intervalle specifique'
cr='Survie relative cumulee'
p_star='Survie attendue pour un intervalle specifique'
cp_star='Survie attendue cumulative'
log_risk='log(n_prime-d/2)'
ln_exact='ln(personnes-temps) (utilisant les temps exacts)'
y_exact='Personnes-temps a risque (utilisant les temps exacts)'
e_exact='Deces attendus (utilisant les temps exacts)'
excess='Exces de risque empirique'
se_p='Erreur standard de P'
se_r='Erreur standard de R'
se_cp='Standard error of CP'
se_cr='Erreur standard de CR'
;
run;

proc print data=&grouped noobs label;
title2 'Life table estimates of patient survival';
title3 'The Ederer II method is used to estimate expected survival';
by &vars;
pageby &lastvar;

```

```
var range n d w n_prime p cp p_star cp_star r cr;  
format fu 3.0 n d w 4.0 n_prime 8.1 p cp p_star cp_star r cr se_p  
      se_r se_cp se_cr 8.5;  
label n='N' d='D' w='W';  
run;
```

B) Modélisation de la survie relative

```

/*****
MODELES.SAS

Modeles de regression pour la survie relative ajustes aux donnees
sur les AVC

Eve-Marie Castonguay Mai 2004
*****/

title1 'Cas AVC diagnostiques au Quebec en 1990-1992';

libname survie 'C:\Documents and Settings\emcastonguay\Mes
documents\AVC'; options fmtsearch=(survival) ls=98 nodate
nonumber;

%let individ=survie.individ ;
%let grouped=survie.grouped ;

/*****
Exact survival times - - - Modele 1 Full likelihood approach
(individual level data) (equation 3.2 du memoire)
*****/
proc nlp data=&individ(where=(fu le 5)) cov=2 vardef=n;
  max loglike;
  parms int fu_2-fu_5 female an_diag91 an_diag92 age2-age8 type2 type3;
  theta = int+fu_2*fu2+fu_3*fu3+fu_4*fu4+fu_5*fu5+an_diag91*an91+an_diag92*an92
    +age2*age_gr2+age3*age_gr3+age4*age_gr4+age5*age_gr5+age6*age_gr6+age7*age_gr7
    +age8*age_gr8+female*sex2+type2*type_autre+type3*type_hemo;
  loglike = d*log(-log(prob)+exp(theta))-exp(theta)*risk;
run;

title; footnote;

/*****
Temps de survie exacts - - - Modele 2 Erreur de structure Poisson
(equation 3.3 du memoire)
*****/
ods output parameterestimates=parmest /* estimation des parametres
*/

  modelinfo=modelinfo          /* Information sur le modele */
  modelfit=modelfit            /* Information sur l'ajustement du modele */
  convergencestatus=converge /* Convergence du modele */
  type3=type3estimates;        /* Estimations de Type III */

```

```

proc genmod data=&grouped(where=(fu le 5)) order=formatted; title2
"Modele d'erreur Poisson (base sur des temps de survie exacts)";
title3 'Effets principaux du modele (suivi, sexe, age, type d'AVC
et annee du diagnostic)';
  fwdlink link = log(_MEAN_-e_exact);
  invlink ilink= exp(_XBETA_)+e_exact;
  class fu sex an_diag age_diag type_avc;
  model d = fu sex an_diag age_diag type_avc / error=poisson offset=ln_exact type3;
  format fu fu. sex sex. age_diag age_diag. an_diag an_diag.;
  output out=test xbeta=xb stdxbeta=stdxb;
run;

ods output close;

data parmest;
  set parmest;
  if df gt 0 then do;
    or=exp(estimate);
    low_rr=exp(estimate-1.96*stderr);
    hi_rr=exp(estimate+1.96*stderr);
  end;
run;

proc print data=parmest label noobs; title4 "Estimation pour beta
et l'excès de risque relatif (ERR=exp(beta))";
  id parameter; by parameter notsorted;
  var level1 estimate stderr or low_rr hi_rr;
  format estimate stderr or low_rr hi_rr 6.3;
  label
    parameter='Parametre'
    level1='Niveau'
    estimate='Estimation'
    stderr='Erreur Standard'
    or='ERR estime'
    low_rr='Limite inferieure IC 95%'
    hi_rr='Limite superieure IC 95%';
run;

/*****
Temps de survie groupes - - Modele 3 Erreur de structure Poisson
(equation 3.8 du memoire)
*****/
ods output parameterestimates=parmest /* Estimation des parametres
*/
  modelinfo=modelinfo      /* Information sur le modele */
  modelfit=modelfit        /* Information sur l'ajustement du modele */
  convergencestatus=converge /* Convergence du modele */
  type3=type3estimates;    /* Estimations de Type III */

```

```

proc genmod data=&grouped(where=(fu le 5)) order=formatted; title2
'Modele d'erreur (approximation des personnes-temps)'; title3
'Modele avec effets principaux (suivi, sexe, age, type d'AVC et
annee de diagnostique)';
  fwdlink link = log(_MEAN_-e);
  invlink ilink= exp(_XBETA_)+e;
  class fu sex age_diag an_diag type_avc;
  model d = fu sex an_diag age_diag type_avc/error=poisson
          offset=log_personnes_temps type3;
  format fu fu. sex sex. age_diag age_diag. an_diag an_diag.;
run;

ods output close;

data parmest;
  set parmest;
  if df gt 0 then do;
    or=exp(estimate);
    low_rr=exp(estimate-1.96*stderr);
    hi_rr=exp(estimate+1.96*stderr);
  end;
run;

proc print data=parmest label noobs; title4 "Estimation pour beta
et l'exces de risque relatif (ERR=exp(beta))";
  id parameter; by parameter notsorted;
  var level1 estimate stderr or low_rr hi_rr;
  format estimate stderr or low_rr hi_rr 6.3;
  label
    parameter='Parametre'
    level1='Niveau'
    estimate='Estimation'
    stderr='Erreur Standard'
    or='ERR estime'
    low_rr='Limite inferieure IC 95%'
    hi_rr='Limite superieure IC 95%';
run;

/*****
Temps de survie groupés - - - Modele 4 Erreur de structure
binomiale (Hakulinen-Tenkanen) (equation 3.10 du memoire)
*****/
ods output parameterestimates=parmest /* estimations des
parametres */
  modelinfo=modelinfo /* Information sur le modele */
  modelfit=modelfit /* Information sur l'ajustement du modele */
  convergencestatus=converge /* Convergence du modele */
  type3=type3estimates; /* Estimations de Type III */

```

```

proc genmod data=&grouped(when=(fu le 5)) order=formatted; title2
"Modele d'erreur binomiale ajusté a des données groupées"; title3
'Modele avec effets principaux (suivi, sexe, age, type d'AVC et
annee de diagnostic)';
  fwdlink link = log(-log(_mean_/p_star));
  invlink ilink = exp(-exp(_xbeta_))*p_star;
  class fu sex an_diag age_diag type_avc;
  model ns/n_prime = fu sex an_diag age_diag type_avc / error=bin type3;
  format fu fu. sex sex. age_diag age_diag. an_diag an_diag.;
run;

ods output close;

data parmest;
  set parmest;
  if df gt 0 then do;
    or=exp(estimate);
    low_rr=exp(estimate-1.96*stderr);
    hi_rr=exp(estimate+1.96*stderr);
  end;
run;

proc print data=parmest label noobs; title4 'Estimates for beta
and relative excess risks (RER=exp(beta))';
  id parameter; by parameter notsorted;
  var level1 estimate stderr or low_rr hi_rr;
  format estimate stderr or low_rr hi_rr 6.3;
  label
    parameter='Parametre'
    level1='Niveau'
    estimate='Estimation'
    stderr='Erreur Standard'
    or='RER estime'
    low_rr='Limite inferieure IC 95%'
    hi_rr='Limite superieure IC 95%';
run;

/*****
/*****
/*****
/***** Modele final choisi *****/

ods output parameterestimates=parmest /* Estimation des parametres
*/
  modelinfo=modelinfo /* Model information */
  modelfit=modelfit /* Information sur le modele */
  convergencestatus=converge /* Convergence du modele */
  type3=type3estimates; /* Estimations de Type III */

```

```

proc genmod data=&grouped(where=(fu le 5)) order=formatted; title2
'Modele d'erreur Poisson ';
  fwdlink link = log(_MEAN_-e);
  invlink ilink= exp(_XBETA_)+e;
  class fu age_diag an_diag type_avc;
  model d = fu an_diag age_diag type_avc fu*age_diag fu*type_avc
          age_diag*type_avc / error=poisson offset=log_personnes_temps type3;
  format fu fu. age_diag age_diag. an_diag an_diag.;
run;

ods output close;

data parmest;
  set parmest;
  if df gt 0 then do;
    or=exp(estimate);
    low_rr=exp(estimate-1.96*stderr);
    hi_rr=exp(estimate+1.96*stderr);
  end;
run;

proc print data=parmest label noobs; title4 "Estimation pour beta
et l'exces de risque relatif (ERR=exp(beta))";
  id parameter; by parameter notsorted;
  var level1 estimate stderr or low_rr hi_rr;
  format estimate stderr or low_rr hi_rr 6.3;
  label
    parameter='Parametre'
    level1='Niveau'
    estimate='Estimation'
    stderr='Erreur Standard'
    or='ERR estime'
    low_rr='Limite inferieure IC 95%'
    hi_rr='Limite superieure IC 95%';
run;

```