ÉCONOMÉTRIE DES DURÉES DE SURVIE
Notes de Cours
MASTER 2 ESA
voies professionnelle et recherche

Gilbert Colletaz

7 octobre 2020

Avertissements

Ce document constitue le support du cours consacré au traitement des données de survie. Actuellement le nombre d'exemples présentés est réduit au minimum, ceux-ci étant réalisés précisément pendant le cours. Compte-tenu de son volume horaire (24 heures), le détail des calculs ainsi que les démonstrations sont souvent seulement esquissés. Celles-ci peuvent être trouvées dans des ouvrages de référence tels que *Statistical Models and Methods for Lifetime Data* de Lawless, *The Statistical Analysis of Failure Time Data* de Kalbfleisch et Prentice, *The Econometric Analysis of Transition Data* de Lancaster, *Survival Analysis : Techniques for Censored and Truncated Data* de Klein et Moeschberger, ou encore *Applied Survival Analysis* de Hosmer et Lemeshow. En français vous avez l'ouvrage récent d'Emmanuel Duguet, *Économétrie appliquée aux variables de durée sous SAS et R*, Economica, 2018. Normalement tout étudiant de deuxième année du Master possède les éléments lui permettant, en cas de besoin, d'être en mesure de les comprendre. Deux ouvrages spécifiquement dédiés à l'utilisation de SAS pour l'analyse des données de survie peuvent également s'avérer utiles : *Survival Analysis Using the Sas System : A Practical Guide* de Allison, et *Survival Analysis Techniques for Medical Research* de Cantor. Enfin on peut trouver sur Internet beaucoup de pages utiles. Voyez notamment celles offertes par UCLA disponibles à partir de l'adresse suivante : *http* : *//www.ats.ucla.edu/stat/sas/seminars/sas_survival/default.htm*.

Enfin, si depuis SAS9 il est possible de réaliser des estimations bayésiennes des modèles de survie, ce type d'estimation sera totalement ignoré dans ce cours pour une raison simple, qui ne préjuge pas de son intérêt. Simplement il n'y a actuellement pas de cours d'économétrie bayésienne dans le cursus du Master, ce qui interdit évidemment d'aborder ces aspects dans les 24 heures imparties.

Pré-requis : Une bonne dose de logique, vos connaissance en statistique, théorie des tests, estimateurs du maximum de vraisemblance et plus particulièrement :

- cours d'économétrie des variables qualitatives (notamment pour la construction des vraisemblances)
- cours d'économétrie non paramétrique, pour l'estimation kernel du risque
- cours de statistique non paramétrique, pour la compréhension de certains tests particulièrement sur l'égalité des courbes de survie

Normalement ce cours a bénéficié des corrections et remarques des étudiants l'ayant suivi. En conséquence, toutes les erreurs restantes leurs sont imputables. N'hésitez cependant pas à me faire part de celles que vous remarqueriez.

Table des matières

1 Introduction **9**
- 1.1 La nature des données de survie . . . 9
- 1.2 La description de la distribution des temps de survie . . . 11
  - 1.2.1 En temps continu . . . 11
  - 1.2.2 En temps discret . . . 13

2 L'approche non paramétrique **15**
- 2.1 L'estimateur de Kaplan-Meier de la fonction de survie : une présentation heuristique 16
- 2.2 Kaplan-Meier comme estimateur du maximum de vraisemblance non paramétrique 20
- 2.3 Les principales hypothèses et leur signification . . . 23
  - 2.3.1 L'hypothèse de censure non informative . . . 23
  - 2.3.2 L'hypothèse d'homogénéité de la population étudiée . . . 24
- 2.4 La variance de l'estimateur de Kaplan-Meier . . . 25
- 2.5 La construction d'IC sur la survie . . . 27
  - 2.5.1 Les intervalles de confiance ponctuels . . . 27
  - 2.5.2 Les bandes de confiance . . . 27
- 2.6 L'estimation de la fonction de risque cumulé . . . 29
- 2.7 L'estimation kernel du risque instantané . . . 30
  - 2.7.1 Le choix de la fonction Kernel . . . 31
  - 2.7.2 Le choix du paramètre de lissage . . . 32
- 2.8 Comparaison de courbes de survie estimées par Kaplan-Meier . . . 34
  - 2.8.1 La statistique du LogRank . . . 34
  - 2.8.2 Le test de Wilcoxon (ou de Gehan) et les autres statistiques pondérées . . . 38
  - 2.8.3 Les tests stratifiés de comparaison des survies . . . 43
  - 2.8.4 Tests d'association entre une variable continue et la survie . . . 45
- 2.9 Un exemple de fonction de risque en temps discret : étude du comportement de réapprovisionnement . . . 48
- 2.10 Les tables de survie - La méthode actuarielle . . . 50
- 2.11 PROC LIFETEST . . . 53

3 L'approche paramétrique **59**
- 3.1 Les modèles AFT et les modèles PH . . . 60
  - 3.1.1 Les Modèles à temps de vie accélérée . . . 60
  - 3.1.2 Les Modèles à risques proportionnels . . . 63
- 3.2 Les principales modélisations AFT . . . 65

3.2.1 La distribution exponentielle . . . 65
3.2.2 La distribution de Weibull . . . 66
3.2.3 La distribution log-normale . . . 66
3.2.4 La distribution log-logistique . . . 68
3.2.5 La distribution Gamma généralisée . . . 68
3.3 Estimation avec différents types de censure et tests sur les coefficients . . . 69
3.4 Choix d'une distribution et tests de spécification . . . 70
3.4.1 Sélection au moyen du test de rapport de vraisemblance . . . 70
3.4.2 Les aides graphiques . . . 71
3.5 estimation de fractiles sur les durées d'événement . . . 77
3.6 Données censurées à gauche, à droite et par intervalle . . . 77
3.6.1 La structuration des données . . . 77
3.6.2 Estimation d'un modèle Tobit via LIFEREG . . . 78
3.7 PROC LIFEREG . . . 82

4 L'approche semi-paramétrique 87
4.1 Le modèle de Cox et son estimation . . . 88
4.1.1 La fonction de vraisemblance partielle . . . 88
4.1.2 La correction de Firth en cas de monotonicité de PL . . . 91
4.1.3 La prise en compte d'événements simultanés . . . 92
4.1.4 Spécification de l'équation à estimer, commandes `Model` et `Class` . . . 95
4.2 Les ratios de risque . . . 99
4.2.1 Interprétation des coefficients et Ratios de Risque . . . 99
4.2.2 Commandes Hazardratio et Contrast . . . 100
4.2.3 Des exemples de sorties . . . 105
4.3 L'estimation de la survie de base . . . 107
4.4 L'analyse stratifiée avec le modèle de Cox . . . 110
4.5 Explicatives non constantes dans le temps . . . 114
4.5.1 Données entrées selon un processus de comptage . . . 115
4.5.2 Explicatives non constantes créées par programme . . . 115
4.6 Tests de validation . . . 116
4.6.1 La qualité de l'ajustement . . . 117
4.6.2 Etude de spécification : résidus de martingales, régression locale et sommes partielles cumulées . . . 117
4.6.3 Repérage des outliers : résidus de déviance, statistiques DFBETA et LD. . . . 122
4.6.4 Tests de l'hypothèse PH - Introduction d'interactions avec le temps, Résidus de Schoenfeld et sommes de transformées de résidus de martingale . . . 125
4.7 La sélection automatique des variables explicatives . . . 129

5 l'ajustement du risque en temps discret 133
5.1 L'écriture du modèle . . . 133
5.1.1 la régression logistique . . . 133
5.1.2 la fonction de lien log-log ou complementary log-log . . . 134
5.2 Son estimation . . . 135
5.2.1 Quels écarts-types utiliser ? . . . 138

6 Suppléments **141**
6.1 Statistiques complémentaires ou alternatives aux ratios de risque 141
6.1.1 Mean survival time, restricted mean survival time, restricted mean time loss 141
6.1.2 Survie médiane 147
6.1.3 Survie conditionnelle 148
6.1.4 Un exemple 149
6.2 L'ajustement de rmst en présence de variables explicatives 153
6.2.1 La régression sur pseudo-observations 154
6.2.2 La régression pondérée par les probabilités inverses 160
6.3 Estimation non paramétrique avec censure par intervalle 162
6.3.1 Les intervalles de Turnbull 163
6.3.2 L'estimation de la survie 164
6.3.3 Intervalles de confiance ponctuels sur la survie 166
6.3.4 Comparaison de courbes de survie sur données censurées 167

Chapitre 1

Introduction

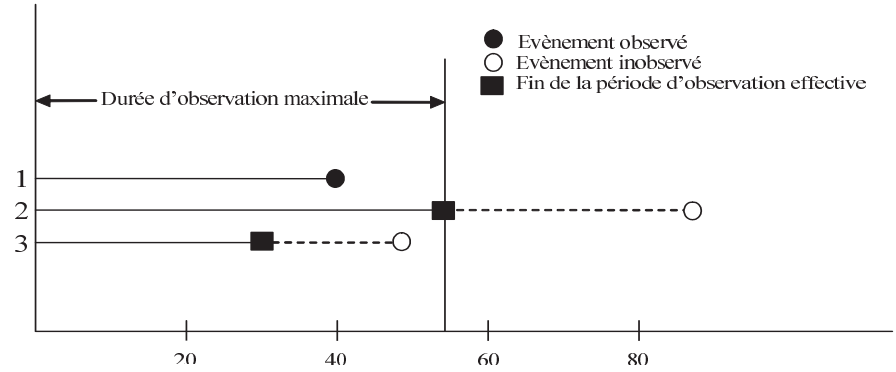
Ce cours a pour objectif la présentation des principales techniques statistiques utilisées pour l'analyse des durées de réalisation d'un ou de plusieurs événements d'intérêt. Le prototype d'événement en question est la mort, d'où le nom le plus courant donné à ces méthodes. Elles s'appliquent cependant à d'autres sortes d'événements (mariage, divorce, rupture d'une relation client, chômage,...). Ces techniques statistiques sont souvent qualifiées d'analyse des biographies ou d'analyse des événements du parcours de vie lorsque les événements analysés découlent d'actions humaines individuelles et d'analyse d'histoire des événements lorsqu'ils résultent d'actions collectives.

Dans cette introduction nous présentons les caractéristiques essentielles des données à analyser ainsi que les outils techniques permettant de décrire leur distribution.

1.1 La nature des données de survie

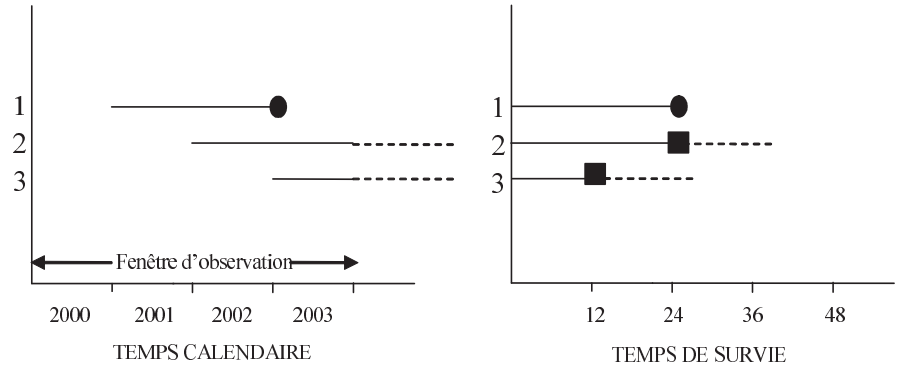
Les temps de survie mesurée à partir d'une origine appropriée ont deux caractéristiques. La première est qu'ils sont non négatifs et tels qu'une hypothèse de normalité n'est généralement pas raisonnable en raison d'une asymétrie prononcée. La seconde est structurelle et tient au fait que pour certains individus l'événement étudié ne se produit pas pendant la période d'observation et en conséquence certaines données sont censurées. Cette censure à droite est la plus courante mais n'est pas la seule censure que l'on peut rencontrer avec des données de survie.

Considérons une étude relative à la durée de survie de patients soumis à un traitement particulier. L'événement d'intérêt est la mort de la personne. Tous les individus sont suivis pendant les 52 semaines suivant la première administration du traitement. On considère plus particulièrement 3 sujets qui vont permettre d'illustrer certaines des caractéristiques les plus fréquentes des données de survie et notamment deux cas possibles de censure à droite.



— L'individu 1 est décédé 40 semaines après le début du traitement. Il s'agit d'une observation non censurée.
— La deuxième personne est toujours vivante au terme des 52 semaines d'observation. Elle décédera après 90 semaines mais cette information n'est pas connue lorsque la constitution de la base de données est arrêtée. Même incomplète l'information est utile puisque l'on sait que le temps de survie réel est supérieur à 52 semaines. Il ne faut donc pas l'éliminer de la base sous peine par exemple de biaiser vers le bas l'estimation de la durée moyenne de survie. Il s'agit d'une censure déterministe car elle ne dépend pas de l'individu considéré mais des conditions de l'expérimentation.
— La troisième personne décède après 50 semaines mais cet évènement n'est pas enregistré dans la base de données car le patient concerné n'a pu être effectivement suivi que pendant 30 semaines. C'est un exemple de censure aléatoire car elle échappe au contrôle de l'expérimentateur. Là encore l'information est incomplète mais non nulle. Par exemple savoir que cet individu a survécu au moins 30 semaines est pertinent pour l'estimation du taux de survie à 20 semaines.

Dans beaucoup d'études l'entrée des individus s'effectue à des temps calendaires différents. Supposons que l'on analyse la durée de l'abonnement à un service, l'évènement d'intérêt étant le non renouvellement du contrat. La fenêtre d'observation s'étend de janvier 2000 à janvier 2004. Une personne ayant souscrit en janvier 2001 et résilié en janvier 2003 a une durée de survie non censurée de 25 mois. Deux nouveaux clients depuis janvier 2002 pour l'un et janvier 2003 pour l'autre et qui le sont toujours en janvier 2004 auront des durées de survie correspondant à une censure déterministe de respectivement 25 mois et 13 mois.



Il s'agit d'étudier la durée passée dans un état préalable à la réalisation, observée ou non, d'un

évènement mais aussi la probabilité de transition d'une situation à une autre. Pour cela il est donc impératif que les individus constituant la base de données soient tous soumis au risque de survenu de l'évènement étudié. Par exemple dans une étude sur la durée et la sortie du chômage seuls des chômeurs ou d'anciens chômeurs seront pris en compte.

1.2 La description de la distribution des temps de survie

L'essentiel de ce cours suppose que les durées de vie sont des variables aléatoires continues. Nous commencerons donc par la présentation des outils de description adaptées aux variables continues. Dans un second temps nous considérerons un temps de survenue des événements de nature discrète. La compréhension des premiers doit faciliter celles des seconds, mais il faut cependant faire attention aux différences d'interprétation qui peuvent apparaître, cela particulièrement sur la fonction de risque.

1.2.1 En temps continu

Pour des données continues la durée de vie T, c'est-à-dire la durée observée d'un individu dans un état initial, est une variable aléatoire définie sur $[0, +\infty[$ de fonction de répartition F. La fonction de survie est définie comme :

$$S(t) = Prob[T > t] = 1 - F(t),\ t \geq 0 \tag{1.1}$$

C'est donc une fonction continue monotone non croissante telle que $S(0) = 1$ et $\lim_{t\to\infty} S(t) = 0$.

Pour décrire cette distribution on peut également recourir à la fonction de densité :

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{1.2}$$

Un concept important dans ces analyses est celui de risque. Considérons la quantité $Prob[t \leq T < t + \Delta t | T \geq t]$. C'est la probabilité de survenue de l'évènement durant l'intervalle de temps $[t + \Delta t[$ sachant qu'il ne s'était pas réalisé avant t. Naturellement, si l'intervalle de temps en question tend vers zéro alors avec une aléatoire continue la probabilité en question tend aussi vers zéro. Les choses changent si on la norme par la durée de l'intervalle lui-même :

$$\frac{Prob[t \leq T < t + \Delta t | T \geq t]}{\Delta t}$$

On passe alors à une évaluation du risque de connaître l'évènement durant l'intervalle de temps considéré. La quantité obtenue mesure en effet le nombre moyen d'évènements que connaîtrait l'individu concerné au cours d'une unité de temps choisie (mois, année par exemple) si les conditions prévalant durant l'intervalle de temps considéré restaient inchangées tout au long de l'unité de temps choisie et pas seulement sur l'intervalle. Par exemple si l'unité est l'année, si la durée de l'intervalle Δt correspond à un mois et si la probabilité de connaître l'évènement au cours de ce mois est de 20% alors l'expression ci-dessus vaut :

$$\frac{Prob[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \frac{20\%}{1/12} = 2.4$$

ce qui signifie qu'en moyenne si les conditions observées pendant le mois en question se maintenaient toute l'année, l'individu connaîtrait en moyenne 2.4 évènements par an ce qui est bien l'évaluation d'un risque. Par exemple un risque de 4 d'attraper un rhume signifie qu'en moyenne sur l'année on contracte 4 rhumes. Ce concept peut paraître absurde pour les évènements qui ne peuvent être répétés tel que le décès. Il est toutefois toujours intéressant de considérer l'inverse du risque qui est une évaluation de la durée moyenne d'attente de la réalisation de l'évènement. Par exemple si l'unité de durée est l'année et qu'au cours d'une expédition lointaine d'un mois on affirme avoir connu une probabilité de mourir de 33%, ce qui correspond à un risque de mourir égal à 4 (évènement qui s'il se produit une fois interdit naturellement la possibilité des 3 autres occurrences) cela implique qu'en moyenne, à conditions inchangées, on peut s'attendre à vivre encore 3 mois.
On va définir la fonction de risque (hasard function) qui apparaît comme une mesure du risque instantané. Attention dans ce cadre d'un temps continu ce n'est pas une probabilité, elle renvoie un réel positif. En particulier, ainsi qu'on vient de le voir, elle peut prendre des valeurs supérieures à l'unité. Sa définition est donc :

$$h(t) = \lim_{\Delta t \to 0} \frac{Prob[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{1.3}$$

Cette fonction est liée aux précédents objets puisqu'en effet avec le théorème des probabilités conditionnelles il vient immédiatement :

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

Il est alors possible de définir le risque cumulé $H(t)$ selon :

$$H(t) = \int_0^t h(s)ds \tag{1.4}$$

Avec l'égalité suivante entre fonction de survie et fonction de risque cumulé :

$$H(t) = -\log[S(t)]$$

en effet

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \log S_t}{dt} \Rightarrow S(t) = \exp[-\int_0^t h(s)ds] = \exp[-H(t)]$$

Toutes ces fonctions sont donc liées entre elles : la connaissance de $S(t)$ permet celle de $f(t)$ via (1.2) et donc celles de $h(t)$ par (1.3) et $H(t)$ par (1.4). De même, la connaissance de $h(t)$ permet celle de $H(t)$ donc de $S(t)$ et finalement de $f(t)$. En d'autres termes, si on se donne une seule de ces fonctions, alors les autres sont dans le même temps également définies. En particulier, un choix de spécification sur la fonction de risque instantané implique la sélection d'une certaine distribution des données de survie.

1.2.2 En temps discret

Ici $t = 1, 2, \ldots, t_k$ et la densité des durées est donnée par un ensemble de probabilités : $f(1) = Pr[T = 1], f(2) = Pr[T = 2], \ldots, f(t_k) = Pr[T = t_k]$.

La survie à la durée t_i étant la probabilité que la réalisation de T soit postérieure à t_i, elle est donnée par :

$$S(t_i) = Pr[T > t_i] = \sum_{t_j > t_i} f(t_j) \tag{1.5}$$

et le risque instantané, toujours défini comme la probabilité de connaître l'événement à un temps quelconque, ici t_i sachant qu'on ne l'a pas connu jusqu'alors, est égal à :

$$h(t_i) = Pr[T = t_i[T \geq t_i] = \frac{f(t_i)}{S(t_i)} \tag{1.6}$$

Notez que, contrairement au cas du temps continu où le risque instantané est positif et peut être supérieur à l'unité, il est en temps discret une probabilité conditionnelle et doit donc être compris entre 0 et 1. C'est une différence entre les deux analyses qu'il faut retenir.

Chapitre 2

L'approche non paramétrique

L'estimateur de la fonction de survie le plus utilisé lorsqu'aucune hypothèse ne veut être faite sur la distribution des temps de survie est l'estimateur de Kaplan-Meier. Dans un premier temps nous en donnons une dérivation heuristique. Dans un second temps nous le présentons comme estimateur du maximum de vraisemblance non paramétrique. Ce faisant nous aurons alors un cadre cohérent d'analyse au sein duquel pourrons être discutées les principales hypothèses nécessaires à sa dérivation et leur signification. Par ailleurs nous serons également en mesure de dériver sa variance, et donc d'apprécier la précision avec laquelle la survie est estimée.
La conséquence immédiate des acquis précédents sera la possibilité de construire des intervalles de confiance ponctuels autour de la survie estimée. Nous présenterons également la construction de bandes de confiance afférente à un seuil de confiance fixé a priori [1].
Si Kaplan-Meier est utile pour estimer une fonction de survie, on peut être intéressé par l'estimation d'autres fonctions qui caractérisent la distribution des temps d'évènements. Nous traiterons donc de l'estimation de la fonction de risque cumulée, avec l'estimateur de Nelson-Aalen. Enfin nous verrons la construction d'estimateurs à noyaux de la fonction de risque cumulée [2].
En pratique, au-delà de la mise en évidence des caractéristiques de la distribution des temps de survie au sein d'une population donnée, il n'est pas rare de s'interroger sur d'éventuelles écarts entre les distributions afférentes à deux ou plusieurs sous-population (par exemple hommes versus femmes, mariés versus célibataires, individus soumis à un traitement particulier versus individus non soumis à ce traitement, etc...). Outre son intérêt propre, la mise en évidence de caractéristiques responsables d'écarts significatifs dans les survies est souvent utilisée comme une première étape de sélection d'explicatives avant la mise en oeuvre d'estimations de modèles paramétriques ou semi-paramétriques. La section suivante présente donc les tests les plus usités d'égalité des fonctions de survie estimées par Kaplan-Meier avec notamment les versions stratifiées de ces tests.
La section suivante est consacrée à l'estimation non paramétrique des caractéristiques de la distribution des temps d'évènements à partir de données issues de tables de survie. Enfin nous présentons, dans la dernière section, les principales commandes et options de la proc LIFETEST de SAS sous SAS 9.2.

1. en faisant toutefois l'impasse sur les démonstrations qui relèvent d'une ré-écriture de KM en termes de processus de comptage permettant de faire appel à la théorie des martingales. A l'évidence le temps imparti à ce cours ne permet pas d'aborder ces aspects.
2. Ces deux estimations sont en maintenant aisément réalisables avec SAS 9.2

2.1 L'estimateur de Kaplan-Meier de la fonction de survie : une présentation heuristique

La fonction de survie est donc définie comme :

$$S(t) = Prob[T > t] = 1 - F(t),\ t \geq 0$$

Soient 2 durées t_1 et t_2 telles que $t_2 > t_1$, alors :

$$Prob[T > t_2] = Prob[T > t_2 \text{ et } T > t_1]$$

puisque pour survivre après t_2 il faut naturellement avoir déjà survécu au moins pendant une durée t_1. . On utilise ensuite le théorème des probabilités conditionnelles et il vient :

$$Prob[T > t_2] = \underbrace{Prob[T > t_2 | T > t_1]}_{(a)} \times \underbrace{Prob[T > t_1]}_{(b)}$$

où

— (a) peut être estimé par $1 - d_{t_2}/n_{t_2}$ où d_{t_2} est le nombre d'individus ayant connu l'évènement en t_2 et n_{t_2} le nombre d'individus qui auraient pu connaître l'évènement en question entre t_1 exclu et t_2. En d'autres termes n_{t_2} est le nombre d'individus à risque au temps t_2. Si le temps était vraiment continu on devrait toujours avoir $d = 1$. . En pratique la périodicité de collecte des données dissimule cette continuité et on observe couramment des valeurs de d supérieures à l'unité traduisant le fait que la discrétisation du temps imposée par le mode de collecte fait que plusieurs individus connaissent l'évènement au *même* instant t. Que ceci ait des conséquences sévères pour l'analyse dépend naturellement du degré de cette agrégation. Il est cependant possible que la nature du problème impose le recours à une analyse en temps discret, par exemple si on étudie l'obtention d'un diplôme avec une succession de jurys semestriels. Dans ce cours nous travaillons essentiellement avec l'hypohèse de distributions continues. Lorsqu'existent des durées censurées entre t_1 inclus et t_2 exclus, la convention retenue est de ne pas prendre en compte les individus concernés dans le calcul de n_{t_2} , nombre d'individus à risque en t_2. Ainsi, si n_{t_1} et n_{t_2} sont respectivement les nombres d'individus à risque en t_1 et t_2, d_{t_1} le nombre d'individus ayant connu l'évènement en t_1 et $c_{[t_1,t_2[}$ le nombre d'individus censurés entre les deux dates, on a $n_{t_2} = n_{t_1} - d_{t_1} - c_{[t_1,t_2[}$.

— (b) est par définition $S(t_1)$.

On a donc :

$$\hat{S}(t_2) = (1 - \frac{d_{t_2}}{n_{t_2}}) \times \hat{S}(t_1) \tag{2.1}$$

L'équation précédente donne une récurrence permettant de calculer $\hat{S}(t)$ pour tout temps d'évènement t observé, sachant qu'initialement $\hat{S}(0) = 1$:

$$\hat{S}(t) = \prod_{i|t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{2.2}$$

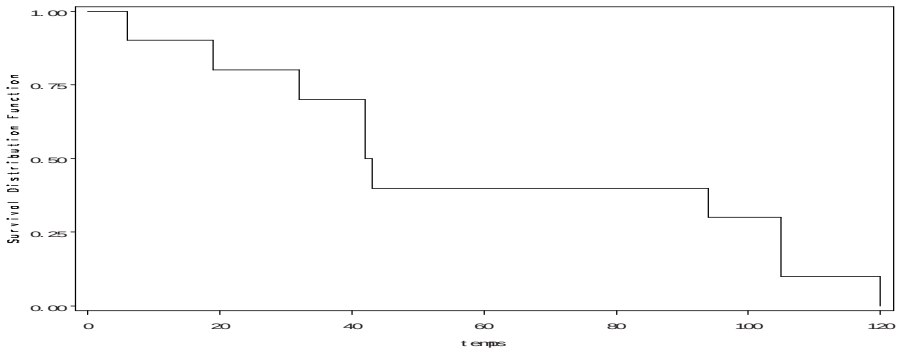
On peut montrer que, sous des conditions assez faibles, l'estimateur de Kaplan-Meier, $\hat{S}(t)$, a asymptotiquement une distribution normale centrée sur $S(t)$. Une de ces conditions est que la censure soit non informative relativement à l'évènement étudié : une façon de comprendre cette condition est que la probabilité de connaître l'évènement étudié à un temps t quelconque est la même pour les individus censurés et les individus non censurés. Cet aspect sera précisé à la section suivante.

Exemple 1 : On ne considère pour l'instant que des données complètes, cad. non censurées relatives à des temps de réalisation d'un évènement mesurés en jours et observé sur 10 individus :

6, 19, 32, 42, 42, 43, 94, 105, 105, 120

t_i	d_i	n_i	$1 - d_i/n_i$	$\hat{S}(t_i)$	$\hat{F}(t_i) = 1 - \hat{S}(t_i)$
0	0	10	1	1	0
6	1	10	1-1/10= 0.90	0.90	0.10
19	1	9	1-1/9 = 0.889	0.80	0.20
32	1	8	1-1/8 = 0.875	0.70	0.30
42	2	7	1-2/7 = 0.7143	0.50	0.50
43	1	5	1-1/5 = 0.80	0.40	0.60
94	1	4	1-1/4 = 0.75	0.30	0.70
105	2	3	1-2/3 = 0.330	0.10	0.90
120	1	1	1-1 = 0	0	1

La représentation graphique associée étant :



— Remarque : $\hat{S}(t)$ est une fonction en escalier dont la valeur change uniquement aux temps correspondant à des évènements observés. En effet à un instant t_i se produit un évènement qui mène à une estimation $\hat{S}(t_i)$. Le prochain évènement se produira à l'instant t_{i+1} et donc entre les temps t_i inclus et t_{i+1} exclus aucune information nouvelle n'apparaît relativement à celle dont on dispose en t_i : il n'y a donc pas lieu de réviser l'estimateur de la fonction de survie.

Exemple 2 : On introduit des données censurées. Dans ce cas la fonction de survie n'est estimée que pour les temps observés mais il faut naturellement ajuster le nombre d'individus à risque. La règle est que pour une durée donnée t_i on ne comptabilise dans les individus risqués que ceux qui ont une date d'évènement égale ou supérieure à t_i ou une durée de censure supérieure à t_i (au passage on notera qu'une convention est que si, pour un individu quelconque, les survenues de l'évènement et de la censure sont concomitantes alors on le considère comme comme non censuré.

En d'autres termes, on impose que la réalisation de l'évènement précède la censure).
Dans la liste ci-dessous relatives à 19 durées mesurées en jours, les données censurées sont signalées par l'exposant * :

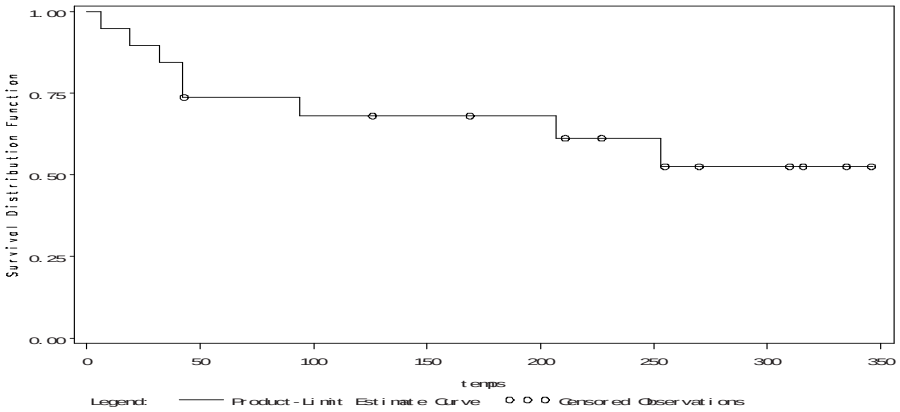
6, 19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316*, 335*, 346*
On obtient alors :

t_i	d_i	n_i	$1 - d_i/n_i$	$\hat{S}(t_i)$	$\hat{F}(t_i) = 1 - \hat{S}(t_i)$
0	0	19	1	1	0
6	1	19	0.947	0.947	0.053
19	1	18	0.944	0.895	0.105
32	1	17	0.941	0.842	0.158
42	2	16	0.875	0.737	0.263
94	1	13	0.923	0.680	0.320
207	1	10	0.90	0.612	0.388
253	1	7	0.957	0.525	0.475

Remarques :

— Il existe des durées supérieures à 253 jours mais elles sont toutes censurées. En conséquence dans ce deuxième exemple, et contrairement au précédent, la valeur estimée de la fonction de survie correspondant au temps d'évènement maximal observé (soit 253 jours) ne s'annule pas. Précédemment nous avions $\hat{S}(120) = 0$ du fait que la durée maximale était non censurée. En d'autres termes, le fait que l'estimateur de la fonction de survie s'annule ne signifie pas que tous les individus ont connu l'évènement étudié mais seulement que la durée maximale ne correspond pas à une censure. Pour vous en persuader, reprenez les chiffres de ce second exemple en remplaçant 346* par 346.

— En ce qui concerne la représentation graphique de la fonction de survie beaucoup vont la tracer jusqu'au temps $t = 253$, ce qui est raisonnable puisque l'estimateur KM n'est pas défini au-delà du temps d'évènement maximal. Toutefois, vous trouverez aussi des présentations qui vont la prolonger jusqu'au temps $t = 346$, maximum des temps censurés qui est ici supérieur au plus grand temps d'évènement connu, avec une horizontale d'ordonnée 0.525. On retrouve une fonction en escalier pour les raisons précédemment avancées : soient t_i et t_{i+1} deux temps d'évènements observés successifs, la révision de l'estimateur calculé en t_i ne se produira qu'en t_{i+1} et les temps censurés compris entre ces deux instants ne seront pris en compte que pour l'évaluation du nombre d'individus à risque en t_{i+1}. Selon ce raisonnement si on a dépassé le maximum des temps d'évènement réalisé alors on doit avoir naturellement une horizontale pour le dernier segment. On obtient ainsi avec ce second exemple le graphe suivant :



Après avoir estimé une fonction de survie il est souvent intéressant d'estimer les quantiles de la distribution des temps de survie. Dans SAS l'estimateur du $p^{\text{ième}}$ centile est donné par $\hat{q}_p = \inf[t|\hat{S}(t) < 1-p]$ ou t est pris dans l'ensemble des temps d'évènement observés. Par défaut le logiciel affiche l'estimation de la médiane ainsi que celles des premier et le troisième quartiles. Ainsi, le premier quartile ($p = 0.25$) est le temps d'évènement au-delà duquel on estime que 75% des individus ne vont pas encore connaître l'évènement : $\hat{q}_{0.25} = \inf[t|\hat{S}(t) < 0.75]$. Dans le premier exemple on a en conséquence $\hat{q}_{0.25} = 32$ jours. De même, l'évaluation du troisième quartile $\hat{q}_{0.75}$ est de 105 jours. Dans le cas où il existe un temps d'évènement t_j tels que $\hat{S}(t_j) = 1-p$ alors le $p^{\text{ième}}$ centile est évalué comme $\hat{q}_p = \frac{1}{2}(t_j + t_{j+1})$. Ainsi, toujours dans le premier exemple, comme $\hat{S}(42) = 0.50$, la médiane est estimée à $\frac{1}{2}(42+43)$, soit 42,5 jours. Dans le deuxième exemple il est impossible d'estimer la médiane et a fortiori le troisième quartile de la distribution puisque l'on n'atteint jamais la valeur de 50%. Ceci se produit en raison de la présence de nombreuses données censurées pour les plus grandes durées d'évènement, ce qui est souvent le cas. Il est possible en revanche d'estimer le premier quartile de la distribution des temps d'évènements : $\hat{q}_{0.25} = 42$ jours : au moins 75% des individus n'ont pas connu l'évènement avant 42 jours. On peut également construire des intervalles de confiance sur ces percentiles et SAS utilise pour cela une méthodologie basée sur un test de signe proposé par Brookmeyer et Crowley (1982), l'intervalle de confiance du $p^{\text{ième}}$ centile IC_p est alors donné par $IC_p = \{t|[1-\hat{S}(t)-p]^2 \leq c_\alpha Var[\hat{S}(t)]\}$, où c_α est la valeur critique au seuil α d'un Chi2 à un degré de liberté, et $Var[\hat{S}(t)]$ est la variance estimée de $\hat{S}(t)$ qui sera définie dans le paragraphe suivant.

Remarques

— La présence de données censurées pour les plus longues durées de vie affecte la qualité de la moyenne empirique en tant qu'estimateur de l'espérance de la distribution des temps de survie. Pour cette raison on préfère utiliser la médiane. Cette préférence est encore renforcée par le fait que très souvent cette distribution est asymétrique. On rappelle d'ailleurs qu'il est toujours utile de commencer une étude par une analyse descriptive simple des séries de travail, ici des temps de survie.

— L'estimateur KM possède une caractéristique contre intuitive : les durées d'évènements longues tendent abaisser la courbe de survie estimée plus que ne le font les évènements observés à durée courte (Cantor et Shuster, 1992, et Oakes, 1993). Ceci provient du fait que n_i décroît avec t_i et donc l'estimation de la probabilité conditionnelle de survenue de l'évè-

nement, à d_i identique, diminue lorsque t_i augmente. Pour vous en persuader, reprenez les données de l'exemple 2 en supposant que l'évènement initialement observé au $6^{\text{ième}}$ jour se soit produit au $230^{\text{ième}}$. La série à considérer est donc :
19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, **230**, 253, 255*, 270*, 310*, 316*, 335*, 346*
Alors que cette substitution correspond à une amélioration de la durée de survie, vous devez vérifier que si initialement une survie d'au moins 253 jours a une probabilité estimée de 0.525, elle est maintenant de 0.500 et est donc, paradoxalement, inférieure.

2.2 Kaplan-Meier comme estimateur du maximum de vraisemblance non paramétrique

L'objectif de cette section est de donner un cadre formel plus assuré permettant la dérivation de l'estimateur de Kaplan-Meier. On va notamment écrire, sans faire d'hypothèse sur la distribution des temps de survie, une fonction de vraisemblance et obtenir KM comme solution au problème de maximisation de cette fonction. Les seules hypothèses sont des conditions de régularité (notamment de continuité et de différentiabilité de la fonction de survie). En outre, cela permettra par la suite d'évaluer la variance de cet estimateur à partir de la matrice d'information de Fisher associée à cette vraisemblance.

On va se situer dans un cadre de censure aléatoire à droite pour lequel un certain nombre de notations et d'hypothèses doivent tout d'abord être précisées. On considère deux aléatoires T_i et C_i qui, pour chaque individu i, donnent respectivement le temps de survenu de l'évènement étudié, t_i et le temps de censure c_i. Dans le cas d'une censure à droite qui nous intéresse ici, le temps de survenue n'est pas toujours connu : ce que l'on observe est la réalisation de T^* définie par $T_i^* = \min(T_i, C_i)$: si $c_i < t_i$ alors le temps de survenu n'est pas connu pour cet individu et le temps t_i^* pris en compte dans les calculs pour cet individu est c_i. Inversement, si $c_i \geq t_i$ alors $t_i^* = t_i$. Au final, $t_i^* = \min(t_i, c_i)$. Cette information est complétée par la valeur d'une indicatrice signalant la présence ou l'absence de la censure : $\delta_i = 1$ si $t_i^* = t_i$ et $\delta_i = 0$ s'il y a censure. On suppose que les T_i sont indépendantes entre elles, que les C_i sont indépendantes entre elles et de plus que le mécanisme de censure est, pour chaque individu, indépendant de la survenue de l'évènement étudié : T_i et C_i sont également indépendantes. On admet enfin que les variables aléatoires $t_i, i = 1, \ldots, n$, ont même distribution avec une fonction de densité $f(t)$ et une fonction de survie $S(t)$. De même les variables C_i ont toutes la même distribution de densité $m(t)$ et de survie $M(t)$.
Ce qui nous intéresse est naturellement d'estimer les caractéristiques de la distribution des temps de réalisation de l'évènement et en particulier la fonction de survie $S(t)$. Pour ce faire, on va s'intéresser à la vraisemblance associée à cette configuration. On commence par écrire la vraisemblance associée à une observation selon qu'elle est censurée ou pas. Il vient :

— Cas d'un temps non censuré : la vraisemblance est donnée par la probabilité de survenue de l'évènement au temps t, soit

$$Pr[T \geq t] - Pr[T > t] = S(t^-) - S(t), \text{où } S(t^-) = \lim_{dt \to 0^-} S(t + dt).$$

Elle est encore égale à $f(t)$ (plus précisément, à $f(t)dt$ si Y est continue).

— Cas d'un temps censuré : la vraisemblance associée est égale simplement égale à la probabilité que T soit supérieure à t, soit $S(t)$.

On note d_i le nombre d'individus qui connaissent l'évènement étudié au temps $t_i, i = 1, \ldots, k$ et m_i le nombre d'individus soumis à une censure sur $[t_i, t_{i+1}[$ à des temps $t_{i,1}, t_{i,2}, \ldots, t_{i,m_i}$. Sur l'échelle des temps, les observations se répartissent donc comme suit :

$$\underbrace{\boxed{0} < t_{0,1} \leq t_{0,2} \leq \ldots \leq t_{0,m_0}}_{m_0 \text{ individus censurés}} < \overbrace{\boxed{t_1}}^{d_1 \text{ non censurés}} \underbrace{\leq t_{1,1} \leq \ldots \leq t_{1,m_1}}_{m_1 \text{ censurés}} < \overbrace{\boxed{t_2}}^{d_2 \text{ non censurés}} \leq \cdots < \overbrace{\boxed{t_k}}^{d_k \text{ non censurés}} \underbrace{\leq t_{k,1} \leq \ldots \leq t_{k,m_k}}_{m_k \text{ censurés}}$$

Au total, la fonction de vraisemblance s'écrit donc :

$$\begin{aligned}
L &= S(t_{0,1})S(t_{0,2}) \cdots S(t_{0,m_0}) \\
&\quad \times \prod_{i=1}^{d_1} [S(t_1^-) - S(t_1)] \\
&\quad \times S(t_{1,1})S(t_{1,2}) \cdots S(t_{1,m_1}) \\
&\quad \times \prod_{i=1}^{d_2} [S(t_2^-) - S(t_2)] \\
&\quad \times \ldots \\
&\quad \times \prod_{i=1}^{d_k} [S(t_k^-) - S(t_k)] \\
&\quad \times S(t_{k,1})S(t_{k,2}) \cdots S(t_{k,m_k}) \\
&= \prod_{i=1}^{m_0} S(t_{0,i}) \prod_{i=1}^{k} \left([S(t_i^-) - S(t_i)]^{d_i} \prod_{j=1}^{m_i} S(t_{i,j}) \right)
\end{aligned}$$

Sachant que l'objectif est de maximiser cette vraisemblance, deux remarques peuvent alors être faites :

— La fonction de survie doit effectuer un saut en $S(t_i)$. En effet, si on pose $S(t_i^-) = S(t_i)$, alors la vraisemblance s'annule. Comme l'objectif est de la maximiser, on doit avoir $S(t_i^-) > S(t_i)$. La fonction de survie étant monotone non croissante, l'écart maximal est obtenu en posant $S(t_i^-) = S(t_{i-1})$.

— Les termes du type $S(t_{i,j}), i = 0, \ldots, k, \quad j = 1, \ldots, m_i$ sont afférents aux individus pour lesquels le temps d'évènement est censuré. Toujours en raison du fait que l'on veut maximiser la vraisemblance et que la fonction de survie $S()$ est monotone non croissante on doit leur donner la plus grande valeur possible. Celle-ci est alors la valeur prise par la fonction sur le temps d'évènement réalisé qui leur est immédiatement antérieur. Ainsi $S(t_{i,j}) = S(t_i)$ avec en particulier $S(t_{0,j}) = S(t_0) = S(0) = 1$.

La fonction de survie est ainsi une fonction étagée avec des sauts aux temps d'évènements non

censurés. Quand à la fonction de vraisemblance, elle s'écrit donc encore :

$$L = \prod_{i=1}^{k} [S(t_{i-1}) - S(t_i)]^{d_i} S(t_i)^{m_i}$$

Enfin, on sait que

$$\begin{aligned} S(t_i) &= Prob[T > t_i] \\ &= Prob[T > t_i, T > t_{i-1}] \\ &= Prob[T > t_i | T > t_{i-1}] \times Prob[T > t_{i-1}] \\ &= Prob[T > t_i | T > t_{i-1}] \times S(t_{i-1}) \end{aligned}$$

Comme $S(t_0) = S(0) = 1$, alors, en notant π_i la probabilité conditionnelle, il vient : $S(t_1) = \pi_1$, $S(t_2) = \pi_1\pi_2$, $S(t_i) = \pi_1\pi_2 \cdots \pi_i$. On peut alors exprimer la vraisemblance en fonction de ces termes π_i. Par la suite, ayant trouvé les estimateurs du maximum de vraisemblance $\hat{\pi}_i$, nous pourrons calculer ceux de $S(t_i)$ grâce aux égalités précédentes et à la propriété d'invariance aux transformations de ces estimateurs. La vraisemblance peut en effet encore s'écrire comme :

$$\begin{aligned} L &= \prod_{i=1}^{k} (\pi_1\pi_2 \cdots \pi_{i-1})^{d_i} (1 - \pi_i)^{d_i} (\pi_1\pi_2 \cdots \pi_i)^{m_i} \\ &= \prod_{i=1}^{k} (1 - \pi_i)^{d_i} \pi_i^{m_i} (\pi_1\pi_2 \cdots \pi_{i-1})^{d_i + m_i} \end{aligned}$$

Finalement en notant n_i le nombre d'individus à risque au temps d'évènement t_i, soit $n_i = \sum_{j \geq i}(d_j + m_j)$, on obtient :

$$L = \prod_{i=1}^{k} (1 - \pi_i)^{d_i} \pi_i^{n_i - d_i}$$

et la log-vraisemblance :

$$\ell = \sum_{i=1}^{k} d_i \log(1 - \pi_i) + (n_i - d_i) \log(\pi_i)$$

Les solutions sont aisément obtenues à partir des conditions du premier ordre :

$$\frac{\delta \ell()}{\delta \pi_i} = -\frac{d_i}{1 - \pi_i} + \frac{n_i - d_i}{\pi_i} = 0 \Rightarrow \hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}, \quad i = 1, \ldots, k$$

L'estimateur de Kaplan-Meier est alors donné par :

$$\hat{S}(t) = \prod_{i | t_i \leq t} (1 - \frac{d_i}{n_i})$$

ce qui est bien l'expression (2.2) déjà vue dans la section précédente page 16.

2.3 Les principales hypothèses et leur signification

A côté des hypothèses de régularité des fonctions manipulées, deux aspects doivent tout particulièrement être pris en considération lors de l'emploi de l'estimateur KM : l'hypothèse de censure non informative d'une part, l'homogénéité de la population étudiée d'autre part.

2.3.1 L'hypothèse de censure non informative

Elle correspond à l'hypothèse d'indépendance entre le processus déterminant le temps de survenue de l'évènement t_i et celui déterminant le temps de censure c_i. Lorsqu'elle n'est pas vérifiée, l'estimateur KM est biaisé. Pour saisir le problème, on peut reprendre la configuration précédente et reconsidérer la construction de la vraisemblance sur les individus plutôt que sur les temps. Dans le cas d'une censure à droite, pour chaque individu i, $i = 1, \ldots, n$, on rappelle que l'on a défini une variable aléatoire T_i^* de réalisation $t_i^* = \min(c_i, t_i)$ où c_i et t_i sont les réalisations de deux variables C_i et T_i de densité et de survie respectives $m(t), M(t)$ et $f(t), S(t)$. On dispose également d'une indicatrice $\delta_i = 1$ si $t_i^* = t_i$ et à 0 sinon. Deux cas sont donc à considérer :

— Celui d'un temps de réalisation d'évènement non censuré. La probabilité associée est :

$$\begin{aligned} Prob\left(T_i^* \in [t_i, t_i^+[, T_i^* = T_i\right) &= Prob\left(T_i \in [t_i, t_i^+[, \delta_i = 1\right) \\ &= Prob\left(T_i \in [t_i, t_i^+[, C_i > t_i\right) \\ &= Prob\left(T_i \in [t_i, t_i^+[\right) \times Prob\left(C_i > t_i\right) \\ &= [f(t_i)dt][M(t_i)] \end{aligned}$$

et donc grâce à l'hypothèse d'indépendance de T_i et C_i, $f_{T_i^*,\delta_i}(t_i, 1) = \frac{Prob\left(T_i^* \in [t_i, t_i^+[, T_i^* = T_i\right)}{dt} = f(t_i)M(t_i)$.

— Cas d'un temps de réalisation d'évènement censuré. La probabilité associée est :

$$\begin{aligned} Prob\left(T_i^* \in [t_i, t_i^+[, T_i^* = C_i\right) &= Prob\left(C_i \in [t_i, t_i^+[, \delta_i = 0\right) \\ &= Prob\left(C_i \in [t_i, t_i^+[, T_i > t_i\right) \\ &= Prob\left(C_i \in [t_i, t_i^+[\right) \times Prob\left(T_i > t_i\right) \\ &= [m(t_i)dt][S(t_i)] \end{aligned}$$

d'où $f_{T_i^*,\delta_i}(t_i, 0) = \frac{Prob\left(T_i^* \in [t_i, t_i^+[, T_i^* = C_i\right)}{dt} = m(t_i)S(t_i)$.

La vraisemblance s'exprime alors comme :

$$
\begin{aligned}
L &= \prod_{i=1}^{n} [f(t_i)M(t_i)]^{\delta_i} [m(t_i)S(t_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} [f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}][M(t_i)^{\delta_i} m(t_i)^{1-\delta_i}] \\
&\propto \prod_{i=1}^{n} [f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}]
\end{aligned}
$$

En l'absence d'indépendance la relation de proportionnalité n'est plus valide et on ne peut plus, pour maximiser L, se contenter de considérer seulement les fonctions caractéristiques des temps de survenue des évènements : il faudrait aussi faire intervenir les densité et fonction de survie afférentes au processus de censure. En d'autres termes, choisir $\hat{S}(t)$ de sorte à maximiser la dernière expression comme effectué dans la section précédente n'aurait aucune raison de fournir une estimation satisfaisante de la survie.

2.3.2 L'hypothèse d'homogénéité de la population étudiée

Dans les précédents développements il est admis que le temps de survenue de l'évènement étudié est, pour tous les individus, tiré dans une même distribution. Ainsi, chaque individu possède la même fonction de survie $S()$, ou bien encore de façon équivalente la même fonction de risque instantané, la même fonction de risque cumulée,... Le non respect de cette hypothèse peut avoir des conséquences sévères, et notamment provoquer des erreurs d'interprétation des résultats des estimations. Pour illustrer ce danger, nous reprenons un exemple connu construit à partir de la fonction de risque plutôt qu'avec la fonction de survie. On suppose que pour chaque individu le risque instantané est une constante. Cependant plutôt que d'admettre l'unicité de cette constante entre tous les individus, on pose maintenant qu'existent J groupes dans l'échantillon de travail tels que tous les individus au sein d'un groupe donné se caractérisent par le même risque mais que ce risque diffère entre les groupes. En d'autres termes si on note G_j le $j^{\text{ième}}$ groupe, $j = 1, \ldots, J$, alors, pour deux individus a et b on a :

— Si a et b appartiennent au même groupe G_j alors $h_a(t) = h_b(t)$
— Si $a \in G_i$ et $b \in G_j$ avec $i \neq j$ alors $h_a(t) = \lambda_i \neq h_b(t) = \lambda_j$

Pour simplifier on admet également que les groupes de risque sont ordonnés : $\lambda_1 > \lambda_2 > \cdots > \lambda_{J-1} > \lambda_J$. Que se passe t'il si on estime une fonction de risque avec un tel échantillon en travaillant sous l'hypothèse d'homogénéité ? Il est tout d'abord évident que la fonction de risque de cet échantillon est un mélange des J fonctions de risque afférentes à chacun des groupes. Par ailleurs l'estimateur à un temps d'évènement donné de cette fonction est obtenu en considérant les individus encore à risque à ce temps. Or, lorsque le temps d'évènement augmente la proportion des individus à faible risque doit s'accroître alors que simultanément la proportion des individus à risqué élevé doit diminuer. Ce mouvement n'est sans doute pas uniforme : on peut avoir une faible probabilité de connaître un évènement et cependant le subir rapidement, et inversement une personne peut avoir une forte probabilité pour que se réalise un évènement sans que celui-ci se produise à court terme. Mais en moyenne cette tendance doit être vérifiée. Ainsi le risque estimé

sur l'échantillon complet doit décroître. Pour autant, il faut se garder d'appliquer cette conclusion à chaque individu puisque l'on sait, par construction, que son risque est constant. Concrètement, imaginez que les individus soient des entreprises de création récente et que l'évènement d'intérêt soit la survenue d'une faillite : le fait d'obtenir une décroissance du risque dans l'échantillon ne signifie pas nécessairement que le risque de faillite est élevée dans les premiers mois qui suivent la création d'une entreprise puis diminue si elle a survécu un certain temps puisque l'on obtiendrait la même évolution avec des entreprises à risque constant certaines très risquées mélangées avec des firmes à risque quasi-nul.

En pratique l'homogénéité supposée des individus, si elle n'est pas respectée, conduit à l'estimation d'un mélange de distributions difficilement interprétable et il faut donc tenter de se mettre dans des conditions où elle n'est pas trop invalide. Ceci est par exemple possible dans certaines expériences (pensez par exemple à des échantillons constitués de souris génétiquement identiques, des plants), mais est plus compliqué notamment sur des données d'entreprises, de clients, etc... La solution est de construire des sous-échantillons au sein lesquels elle doit être mieux vérifiée. Ainsi, on pourra distinguer les clients selon leur sexe ou/et leur catégorie d'âge si on pense que la probabilité de survenue de l'évènement diffère entre les hommes et les femmes, les jeunes et les adultes, etc... L'estimation KM est alors réalisée séparément sur chacun de ces sous-échantillons. Nous verrons par la suite qu'il est d'ailleurs possible de tester l'égalité des différentes fonctions de survie, et donc de justifier ou non le découpage de l'échantillon initial. L'autre solution sera d'inclure les caractéristiques en question comme explicatives des paramètres de la fonction de survie et/ou de risque, ce qui sera réalisé par les modélisations paramétrique ou semi-paramétrique dans les chapitres suivants.

2.4 La variance de l'estimateur de Kaplan-Meier

Pour apprécier la précision de l'estimation de $S(t)$ il est utile d'estimer la variance de l'estimateur $\hat{S}(t)$. Pour ceci on peut employer des résultats dérivés dans le cadre de la théorie de l'estimation par du maximum de vraisemblance. On sait notamment que la variance asymptotique de $\hat{\theta}_{MV} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)'$ peut être évaluée par :

$$\hat{V}(\hat{\theta}_{MV}) = -\left(\frac{\delta^2 \ell(\theta)}{\delta\theta_i \delta\theta_j}|_{\theta=\hat{\theta}_{MV}}\right)^{-1}_{i,j=1,\ldots,k}$$

Partant de là, on peut obtenir la formule dite de Greenwood qui est la plus utilisée dans la littérature.

On part de la formule de calcul de l'estimateur KM :

$$\hat{S}(t) = \prod_{i|t_i \leq t}(1 - \frac{d_i}{n_i}) = \prod_{i|t_i \leq t} \hat{\pi}_i$$

et donc :

$$\log(\hat{S}(t)) = \sum_{i|t_i \leq t} \log(1 - \frac{d_i}{n_i}) = \sum_{i|t_i \leq t} \log(\hat{\pi}_i)$$

où

— d_i/n_i est la proportion d'individus ayant connu l'évènement parmi les individus à risque en t_i, et donc
— $1 - d_i/n_i$ est la proportion d'individus n'ayant pas connu l'évènement parmi les individus à risque en t_i.

L'idée est de calculer la variance de la somme précédente, dont on montrera qu'elle est constituée d'éléments orthogonaux entre eux et donc simplement égale à la somme des variances des éléments en question pour, finalement, remonter de la variance de $\log(\hat{S}(t))$ à la variance de $\hat{S}(t)$.

On a immédiatement d'une part :

$$\frac{\delta^2 \ell()}{\delta \pi_i^2} = \frac{n_i - d_i}{\pi_i^2} - \frac{d_i}{(1-\pi_i)^2}, \quad i = 1, \ldots, k$$

et, d'autre part :

$$\frac{\delta^2 \ell()}{\delta \pi_i \delta \pi_j} = 0, \quad i, j = 1, \ldots, k \text{ et } j \neq i$$

La matrice de VarCov est donc diagonale. Évaluée en $\hat{\pi}_i = 1 - \frac{d_i}{n_i}$, on arrive alors $Cov(\hat{\pi}_i, \hat{\pi}_j) = 0$ si $i \neq j$ et

$$Var(\hat{\pi}_i) = \frac{d_i(n_i - d_i)}{n_i^3}$$

En conséquence en utilisant la méthode delta [3],

$$\begin{aligned} Var[\log(\hat{\pi}_i)] &= \hat{\pi}_i^{-2} Var[\hat{\pi}_i] \\ &= \frac{n_i^2}{(n_i - d_i)^2} \times \frac{d_i(n_i - d_i)}{n_i^3} \\ &= \frac{d_i}{(n_i - d_i) n_i} \end{aligned}$$

Et donc :

$$Var[\log(\hat{S}(t))] = \hat{\sigma}^2_{lS_t} = \sum_{i|t_i \leq t} \frac{d_i}{(n_i - d_i) n_i}$$

On veut $Var[\hat{S}(t)]$. Sachant que l'on connaît $Var[\log(\hat{S}(t))]$ et que naturellement $Var[\hat{S}(t)] = Var[\exp\{\log(\hat{S}(t))\}]$, il suffit de reprendre la méthode delta appliquée maintenant à $g() = \exp()$ pour obtenir finalement la formule de Greenwood :

$$\begin{aligned} Var[\hat{S}(t)] &= Var[\exp\{\log(\hat{S}(t))\}] \\ &= [\exp\{\log(\hat{S}(t))\}]^2 \quad Var[\log(\hat{S}(t))] \\ &= \hat{S}(t)^2 \times \sum_{i|t_i \leq t} \frac{d_i}{(n_i - d_i) n_i} \\ &= \hat{S}(t)^2 \times \hat{\sigma}^2_{lS_t} \end{aligned} \tag{2.3}$$

3. Rappel : soit à calculer $Var[g(x)]$ où $g()$ est une fonction continue dérivable. Un développement de Taylor à l'ordre 1 au voisinage de x_0 donne : $g(x) = g(x_0) + (x - x_0)\frac{\partial g(x)}{\partial x}_{|x=x_0} = g(x_0) + (x - x_0) g'_{x_0}$ et donc $Var[g(x)] = g'^2_{x_0} Var(x)$. Ici $g(x) = \log(x)$ et g'(x)=1/x.

2.5 La construction d'IC sur la survie

2.5.1 Les intervalles de confiance ponctuels

Il s'agit de trouver deux bornes b_{L_t} et b_{Ut} telles que $\forall t > 0$ on ait : $Prob[b_{U_t} \geq S(t) \geq b_{L_t}] = 1 - \alpha$, ou α est un seuil de risque fixé a priori.

Le point de départ est relativement complexe à obtenir et sera simplement admis ici : on peut montrer que $\sqrt{n}(\hat{S}(t) - S(t))/S(t)$ converge vers une martingale gaussienne centrée. Une des conséquences est que la distribution asymptotique de $\hat{S}(t)$ est gaussienne et centrée sur $S(t)$. Compte-tenu des résultats précédents, son écart-type estimé, noté $\hat{\sigma}_{S_t}$, est donné par :

$$\hat{\sigma}_{S_t} = \hat{\sigma}_{lS_t}\hat{S}(t), \tag{2.4}$$

et donc un intervalle de confiance au seuil $100(1-\alpha)\%$ peut être construit selon :

$$\hat{S}(t) \pm z_{\alpha/2}\hat{\sigma}_{S_t} \tag{2.5}$$

où $z_{\alpha/2}$ est le fractile de rang $100 \times \alpha/2$ de la distribution normale standardisée.

Un inconvénient de la construction de l'IC avec la formule précédente est que les bornes obtenues peuvent être extérieures à l'intervalle $[0, 1]$. Une solution est de considérer une transformée de $S(t)$ via une fonction $g()$ continue, dérivable et inversible telle que $g(S(t))$ appartienne à un espace plus large idéalement non borné et pouvant mieux approximer une va gaussienne. La méthode delta autorise alors l'estimation de l'écart-type de l'objet ainsi créé au moyen de $\hat{\sigma}_{g(S_t)}$ définit par $\hat{\sigma}_{g(S_t)} = g'(\hat{S}_t)\hat{\sigma}_{S_t}$. L'intervalle de confiance associé au seuil de risque α est construit comme $g^{-1}\left(g(\hat{S}_t) \pm z_{\alpha/2}g'(\hat{S}_t)\hat{\sigma}_{S_t}\right)$. La transformation la plus usitée est $g(S_t) = \log[-log(S_t)]$, et dans ce cas [4] :

$$\hat{\sigma}_{\log[-\log(S_t)]} = \frac{\hat{\sigma}_{S_t}}{\hat{S}_t \log(\hat{S}_t)} \text{ et } \hat{S}_t^{\exp\left(\pm z_{\alpha/2}\frac{\hat{\sigma}_{S_t}}{\hat{S}_t \log(\hat{S}_t)}\right)}.$$

2.5.2 Les bandes de confiance

Il s'agit maintenant de trouver une région du plan qui contienne la fonction de survie avec une probabilité égale à $1 - \alpha$, ou encore un ensemble de bornes b_{L_t} et b_{U_t} qui, avec une probabilité $1 - \alpha$, encadre $S(t)$ pour tout $t \in [t_L, t_U]$. Parmi les solutions proposées, les deux plus couramment employées et disponibles dans SAS 9.2 sont d'une part les bandes de Hall-Wellner et d'autre part les bandes de Nair ("equal precision bands"). Si t_k est le temps d'évènement maximal observé dans l'échantillon, alors pour les bandes de Nair on a les restrictions suivantes $0 < t_L < t_U \leq t_k$, en revanche, avec Hall-Wiener on peut autoriser la nullité de t_l, soit $0 \leq t_L < t_U \leq t_k$. Techniquement l'obtention de ces bandes est complexe [5], et leur utilité pratique par rapport aux intervalles ponctuels n'est pas évidente. En particulier, du fait du caractère joint, pour un t donné leur étendue est

4. On peut également utiliser des transformations de type log, arc-sinus de la racine carrée ou logit dans la plupart des logiciels définies respectivement par $g(S_t) = \log[S_t]$, $g(S_t) = sin^{-1}[\sqrt{S_t}]$, $g(S_t) = \log[S_t/(1 - S_t)]$

5. Le point de départ utilise le fait que $\sqrt{n}\frac{\hat{S}(t)-S(t)}{S(t)}$ converge vers une martingale gaussienne centrée. On passe ensuite par une transformation faisant apparaître un pont brownien $\{W^0(x), x \in [0, 1].\}$, pondéré par $1/\sqrt{x(1-x)}$ chez Nair, permettant de récupérer les valeurs critiques idoines

plus large que celle de l'IC ponctuel correspondant. Dans ce qui suit nous donnons les expressions obtenues en l'absence de transformation. Il peut être encore possible d'appliquer les transformations log, log-log, arc-sinus de la racine carrée ou logistique. En pratique, il est conseillé d'utiliser une transformation avec les bandes de Nair alors que le recours à une transformation serait moins utile sur les bandes de Hall-Wellner.

Les bandes de confiance de Hall-Wellner

Sous l'hypothèse de continuité des fonctions de survie de $S(t)$ et $M(t)$ afférentes respectivement au temps d'évènement et au temps de censure, Hall et Wellner montrent que pour tout $t \in [t_L, t_U]$ l'IC joint au seuil de risque α est donné par :

$$\hat{S}(t) \pm h_\alpha(x_L, x_U) n^{-\frac{1}{2}} [1 + n\hat{\sigma}^2_{lS_t}] \hat{S}(t), \tag{2.6}$$

où x_L et x_U sont donnés par $x_i = n\hat{\sigma}^2_{lS_{t_i}} / (1 + n\hat{\sigma}^2_{lS_{t_i}})$ pour $i = L, U$ et $h_\alpha(x_L, x_U)$ est la borne vérifiant $\alpha = Pr\left[\sup_{x_L \leq x \leq x_U} |W^0(x)| > h_\alpha(x_L, x_U)\right]$.

Les equal precision bands de Nair

L'emploi d'un pont Brownien pondéré va notamment modifier les bornes des IC. Pour tout $t \in [t_L, t_U]$ celles-ci sont alors données par :

$$\hat{S}(t) \pm e_\alpha(x_L, x_U) \hat{\sigma}_{S_t}, \tag{2.7}$$

la borne $e_\alpha(x_L, x_U)$ vérifiant $\alpha = Pr\left[\sup_{x_L \leq x \leq x_U} \frac{|W^0(x)|}{\sqrt{x(1-x)}} > e_\alpha(x_L, x_U)\right]$.

Si on compare (2.5) et (2.7), on voit que les bornes afférentes aux bandes de Nair sont proportionnelles aux bornes des IC ponctuels et correspondent simplement à un ajustement du seuil de risque utilisé dans ces dernières.

Un exemple

Pour illustrer les points précédents, on utilise des données de Klein et moeschberger(1997) distribuées avec l'installation de SAS (fichier BMT).

```
proc format;
value risk 1='ALL' 2='AML-Low Risk' 3='AML-High Risk';
data BMT;
input Group T Status @@;
format Group risk.;
label T='Disease Free Time';
datalines;
1 2081 0 1 1602 0 1 1496 0 1 1462 0 1 1433 0 1 1377 0 1 1330 0 1 996 0 1 226 0 1
1199 0 1 1111 0 1 530 0 1 1182 0 1 1167 0 1 418 1 1 383 1 1 276 1 1 104 1 1 609 1
1 172 1 1 487 1 1 662 1 1 194 1 1 230 1 1 526 1 1 122 1 1 129 1 1 74 1 1 122 1 1
86 1 1 466 1 1 192 1 1 109 1 1 55 1 1 1 1 1 107 1 1 110 1 1 332 1 2 2569 0 2 2506
```

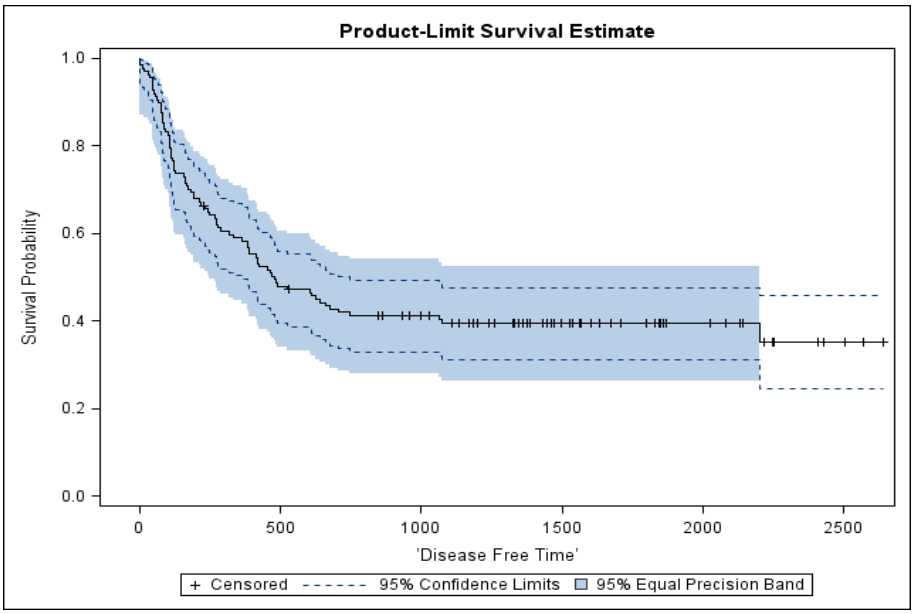


FIGURE 2.1 – Intervalles de confiances ponctuels et bande de confiance

```
0 2 2409 0 2 2218 0 2 1857 0 2 1829 0 2 1562 0 2 1470 0 2 1363 0 2 1030 0 2 860 0
2 1258 0 2 2246 0 2 1870 0 2 1799 0 2 1709 0 2 1674 0 2 1568 0 2 1527 0 2 1324 0
2 957 0 2 932 0 2 847 0 2 848 0 2 1850 0 2 1843 0 2 1535 0 2 1447 0 2 1384 0 2
414 1 2 2204 1 2 1063 1 2 481 1 2 105 1 2 641 1 2 390 1 2 288 1 2 421 1 2 79 1 2
748 1 2 486 1 2 48 1 2 272 1 2 1074 1 2 381 1 2 10 1 2 53 1 2 80 1 2 35 1 2 248 1
2 704 1 2 211 1 2 219 1 2 606 1 3 2640 0 3 2430 0 3 2252 0 3 2140 0 3 2133 0 3
1238 0 3 1631 0 3 2024 0 3 1345 0 3 1136 0 3 845 0 3 422 1 3 162 1 3 84 1 3 100 1
3 2 1 3 47 1 3 242 1 3 456 1 3 268 1 3 318 1 3 32 1 3 467 1 3 47 1 3 390 1 3 183
1 3 105 1 3 115 1 3 164 1 3 93 1 3 120 1 3 80 1 3 677 1 3 64 1 3 168 1 3 74 1 3
16 1 3 157 1 3 625 1 3 48 1 3 273 1 3 63 1 3 76 1 3 113 1 3 363 1;
```

Dans cet exemple on ne considère pas les informations relatives à la variable risk. Les estimations KM de la survie, les intervalles de confiance ponctuels et une bande de confiance, ici de Nair, sont obtenus avec les instructions suivantes :

```
proc lifetest data=BMT plots=s(cl cb=ep); time T * Status(0); run;
```

Le seuil de risque par défaut est utilisé ($\alpha = 5\%$) et les résultats sont présentés dans le graphique (2.1).

2.6 L'estimation de la fonction de risque cumulé

Nelson (1972) et Aalen (1978) ont proposé un estimateur de la fonction de risque cumulée $H(t)$. Connu sous le nom d'estimateur de Nelson-Aalen, il est donné par :

$$\tilde{H}(t) = \sum_{i|t_i \leq t} \frac{d_i}{n_i} \tag{2.8}$$

Un autre estimateur également souvent employé est l'estimateur de Breslow ou de Peterson. Il est obtenu à partir de l'estimateur KM de la survie et reprend l'équation liant les deux fonctions, soit :

$$\hat{H}(t) = -\log \hat{S}(t) \tag{2.9}$$

On peut montrer que $\tilde{H}(t) < \hat{H}(t)$: l'estimateur de Nelson-Aalen est toujours inférieur à l'estimateur de Breslow [6]. Il n'y a cependant aucune raison de privilégier l'un par rapport à l'autre [7].On peut naturellement.

En ce qui concerne la précision de ces estimateurs, on peut estimer la variance de l'estimateur de Peterson par :

$$Var(\hat{H}(t)) = \frac{Var(\hat{S}(t))}{\hat{S}(t)^2} \tag{2.10}$$

où $Var(\hat{S}(t))$ est la variance de l'estimateur KM dérivée précédemment.

Pour l'estimateur de Nelson-Aalen, deux choix asymptotiquement équivalents sont offerts. Soit :

$$Var(\tilde{H}(t)) = \sum_{i|t_i \leq t} \frac{d_i(n_i - d_j)}{n_i^3}, \tag{2.11}$$

soit, et cette deuxième expression est préférable sur petits échantillons :

$$Var(\tilde{H}(t)) = \sum_{i|t_i \leq t} \frac{d_i}{n_i^2}. \tag{2.12}$$

C'est cette dernière formulation qui utilisée dans la proc LIFETEST de SAS)

2.7 L'estimation kernel du risque instantané

Des estimateurs à noyaux de la fonction de risque ont été proposées tant pour des données groupées, du type table de survie que nous étudions dans une section ultérieure, que pour des données individuelles supposant des durées continues. Nous n'aborderons que ce dernier cas qui, actuellement, est le seul implémenté dans SAS.

Par la suite, l'ensemble des calculs ne portent que sur les temps d'évènements observés. L'idée est de proposer un estimateur lissé du risque instantané à partir de l'estimation du risque cumulé. Pour cette dernière on utilise l'estimateur canonique qui est celui de Nelson-Aalen vu au point précédent. Plusieurs méthodes de lissage ont été proposées dont les fonctions splines, mais la plus

6. Ceci vient du fait que la fonction log étant concave elle se situe sous sa tangente et donc, si on considère un développement de Taylor à l'ordre 1, 1l vient $log(1 - x^+) < x^+$. Comme d'une part $\hat{H}(t) = -\sum_{i|t_i \leq t} \log(1 - \frac{d_i}{n_i})$ et $\tilde{H}(t) = \sum_{i|t_i \leq t} \frac{d_i}{n_i}$, on obtient immédiatement la propriété annoncée.

7. Disposant de l'estimateur de Nelson-Aalen du risque cumulé, on peut naturellement remonter vers la survie en exploitant toujours la relation liant les deux fonctions. C'est ce que réalise l'estimateur de Fleming et Harrington : $\tilde{S}(t) = \exp -\tilde{H}(t)$.

usitée (et en tout cas la seule disponible dans SAS) recourt à l'emploi de fonctions de type kernel K associées à un choix de bandwidth b, soit :

$$\hat{h}_n(t) = \int b^{-1} K\left(\frac{t-x}{b}\right) d\tilde{H}_n(t) \tag{2.13}$$

L'absence de biais asymptotique, la normalité asymptotique et la convergence en moyenne quadratique de $\hat{h}_n(t)$ peut être obtenue sous des conditions de régularité [8]

La fonction de risque cumulée $\tilde{H}_n(t)$ étant une fonction à sauts aux temps d'évènements observés t_i, en posant $\triangle\tilde{H}_n(t_i) = \tilde{H}_n(t_i) - \tilde{H}_n(t_{i-1})$ pour $i = 1, 2, \ldots, k$, il vient :

$$\hat{h}_n(t) = \frac{1}{b}\sum_{i=1}^{k} K\left(\frac{t-t_i}{b}\right) \triangle\tilde{H}_n(t_i) \tag{2.14}$$

Quand à sa variance, elle est donnée par :

$$s^2(\hat{h}_n(t)) = \frac{1}{b^2}\sum_{i=1}^{k} K\left(\frac{t-t_i}{b}\right)^2 \triangle Var(\tilde{H}_n(t_i))$$

En conséquence on peut construire des intervalles de confiance ponctuels de la forme $\hat{h}_n(t) \pm z_{\alpha/2} s(\hat{h}_n(t))$, où $z_{\alpha/2}$ est le fractile afférent au seuil de risque α dans la distribution gaussienne standardisée. Comme pour les IC sur la survie, on préfère appliquer une transformation sur $\hat{h}_n(t)$. La proc LIFETEST utilise une transformée logarithmique ce qui conduit à l'IC suivant :

$$\hat{h}_n(t) \exp\left[\pm \frac{z_{\alpha/2} s(\hat{h}_n(t))}{\hat{h}_n(t)}\right]$$

2.7.1 Le choix de la fonction Kernel

Dans SAS 9.2, les 3 choix suivants sont possibles :

— un kernel uniforme : $K_U(x) = \frac{1}{2}$, $-1 \leq x \leq 1$
— un lissage de noyau Epanechnikov : $K_E(x) = \frac{3}{4}(1-x^2)$, $-1 \leq x \leq 1$
— un lissage biweight : $K_{BW}(x) = \frac{15}{16}(1-x^2)^2$, $-1 \leq x \leq 1$

En pratique le lissage Epanechnikov est souvent recommandé même si on admet que le choix du noyau a peu d'impact sur la valeur de l'estimation.

Il faut naturellement se méfier des effets de bord : les méthodes de lissage précédentes deviennent douteuses lorsque le support du kernel dépasse l'étendue des données disponibles c'est à dire au voisinage des temps d'évènements les plus faibles et les plus élevés. Il est alors nécessaire de remplacer les fonctions kernel symétriques par des fonctions asymétriques lorsque $t < b$ d'une part et lorsque $t_k - b \leq t_k$ d'autre part [9]

8. la fonction de hasard est k-fois différentiable (le plus souvent k=2), la fonction kernel est d'ordre k (soit $\int K(x)dx = 1$, $\int K^2(x)dx < \infty$, $\int x^j K(x)dx = 0$ pour $1 < j < k$ et $0 < \int x^k K(x)dx < \infty$), enfin le paramètre de bandwidth vérifie $\lim_{n\to\infty} b_n = 0$ et $\lim_{n\to\infty} nb_n = \infty$.

9. Voir la documentation de proc LIFETEST pour l'expression des fonctions alors mises en oeuvre par SAS.

2.7.2 Le choix du paramètre de lissage

Contrairement au choix du noyau, celui de b est essentiel : en augmentant sa valeur on risque de trop lisser et masquer des caractéristiques pertinentes, en la diminuant on risque de révéler des évolutions très irrégulières essentiellement dues à des bruits aléatoires. Il s'agit en fait d'arbitrer entre le biais et la variance de l'estimateur $\hat{h}_n(t)$ comme le montre les résultats suivants obtenus dans le cadre du modèle à censure aléatoire :

$$\begin{aligned} biais(\hat{h}_n(t)) &= b^k[h^k(t)B_k + 0(1) \\ Var(\hat{h}_n(t)) &= \frac{1}{nb}\left(\frac{h(t)}{[S(t)M(t)]}V + 0(1)\right) \end{aligned}$$

où $B_k = (-1)^k/k! \int x^k K(x)dx$ et $V = \int K^2(x)dx$.

Actuellement SAS utilise une valeur de b qui minimise l'erreur quadratique moyenne intégrée [10]. On notera que la valeur de ce paramètre est donc fixe pour tous les temps d'évènement t_i. On peut mettre en évidence un mauvais comportement des estimations résultantes en raison à nouveau d'un effet de bord notamment pour les t_i élevés : le nombre d'observations diminuant avec t il en résulte une baisse des observations disponibles pour les calculs lorsqu'on travaille avec un paramètre de lissage constant et des t_i de plus en plus grands. Une autre façon de percevoir ce problème est de noter que lorsque le temps d'évènement augmente alors la fonction de survie tend vers zéro et, d'après la dernière équation, la variance de l'estimateur explose. Une solution est de modifier la valeur de b en un $b(t)$ qui soit une fonction croissante de t. Cette possibilité d'emploi d'un paramètre de bandwidth local n'est pas disponible avec SAS 9.22.

Un exemple

On reprend les données du fichier BMT utilisé dans la section précédente. Afin d'estimer le risque ponctuel, nous avons utilisé deux valeurs pour le paramètre bandwith : l'une correspond à sa valeur optimale (94.47), l'autre est imposée à (180.0). Les instructions utilisées sont de la forme :

```
proc lifetest data=BMT plots=h(bw=180);
time T * Status(0);
run;
```

Les résultats sont regroupés dans le graphique 2.1. Par défaut, c'est un Kernel de type Epanechnikov qui est mis en oeuvre. On remarque bien l'effet de lissage accentué associé à l'augmentation du paramètre de bandwith. Par ailleurs, il semble que le risque de décès soit élevé au moment et peu après la date du diagnostic, et qu'il tend ensuite à décroître assez régulièrement pour atteint un plateau où il est pratiquement nul entre 1200 et 2000 jours, avant de remonter vers les plus longues durées. Il faut cependant se souvenir des effets de bords et ne pas commenter une évolution qui serait en fait essentiellement de leur fait. Ces effets affectent la précision des estimations et une façon de les mettre en évidence est de construire des intervalles de confiance autour de la fonction lissée. Cela est effectué dans le graphique 2.2. Dans le présent exercice, on observe des amplitudes pour les IC aux durées faibles et élevées qui interdisent de commenter les évolutions observées sur ces temps. La seule conclusion raisonnable concerne la décroissance du risque quelque mois après le diagnostic.

10. Rappel : $\text{MISE}(\hat{h}_n(t)) = E\left(\int[\hat{h}_n(t) - h(t)]^2 dt\right)$.

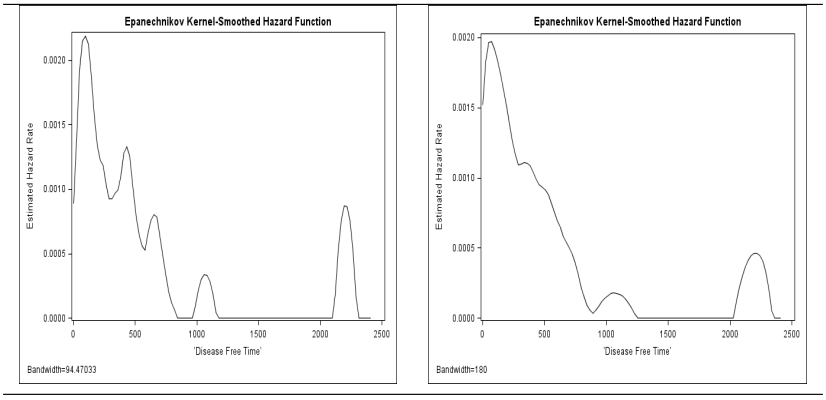


TABLE 2.1 – Estimation kernel de la fonction de risque instantanée

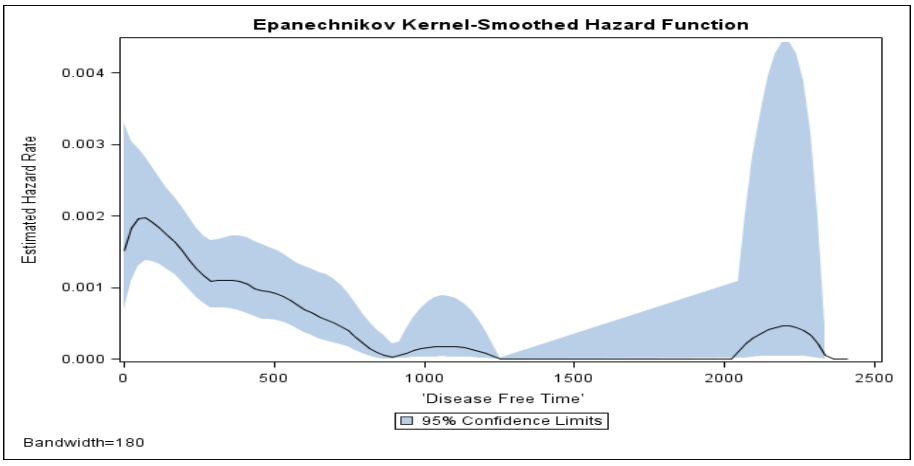


FIGURE 2.2 – Estimation kernel de la fonction de risque instantanée avec intervalles de confiance

2.8 Comparaison de courbes de survie estimées par Kaplan-Meier

Après avoir estimé les courbes de survie sur deux groupes d'individus ou plus, on est souvent amener à vouloir les comparer. L'exemple le plus simple est celui de deux échantillons d'individus issus initialement de la même population mais dont l'un a été soumis à une intervention (par exemple un traitement, une action marketing,...). La question est alors de tester l'efficacité de cette intervention. Il est extrêmement important de s'assurer qu'à l'exception de l'intervention en question, les deux ensembles d'individus possède les mêmes autres caractéristiques. Dans le cas contraire, une divergence des courbes de survie ne peut pas être attribuable à la seule intervention dont on cherche à apprécier l'impact. Par exemple que les caractéristiques d'âges, de sexe, de catégories socioprofessionnelles sont les mêmes dans les deux groupes dont l'un est la cible d'une action marketing et l'autre pas. La plupart des statistiques utilisées sont fondées sur des tableaux de contingences construits sur l'ensemble des temps d'évènements. Ce sont donc des statistiques de rang, la plus couramment utilisée étant la statistique dite du *LogRank* [11].

2.8.1 La statistique du LogRank

Aussi appelée statistique de Mantel-Haenzel, elle doit son nom à Peto et Peto (1972) qui la dérive en considérant les estimateurs du logarithme des fonctions de survie. Pour simplifier nous l'exposerons relativement en détail pour 2 seulement groupes d'individus, avant d'aborder l'extension à plus de deux groupes.

La comparaison de deux fonctions de survie

Soient 2 groupes d'individus indicés 1 et 2 et les effectifs observés au temps t_i :

Groupe	1	2	total
Individus ayant connu l'évènement	d_{1i}	d_{2i}	d_i
Individus n'ayant pas connu l'évènement	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
Individus risqués	n_{1i}	n_{2i}	n_i

L'hypothèse nulle est l'égalité des courbes de survie. Sous H0 la proportion attendue d'évènements à un temps t_i quelconque est donnée par d_i/n_i et le nombre espéré d'évènements au sein des groupes est obtenu en appliquant cette proportion à l'effectif observé de chacun des groupes. Soit :

$$\begin{aligned} e_{1i} &= n_{1i}\frac{d_i}{n_i} \text{ pour le } 1^{\text{er}} \text{ groupe} \\ e_{2i} &= n_{2i}\frac{d_i}{n_i} \\ &= (n_i - n_{1i})\frac{d_i}{n_i} \\ &= d_i - e_{1i} \text{ pour le second groupe} \end{aligned}$$

11. SAS donne également par défaut une statistique de type LRT. Celle-ci suppose une distribution exponentielle des durées qui n'a aucune raison d'être généralement valide. Pour cette raison nous ne la traiterons pas ici. En revanche, elle réapparaîtra dans le chapitre 3 consacré aux modèles paramétriques.

Les règles de construction des données nécessaires à ces calculs sont celles utilisées dans les calculs de la statistique KM. Ainsi, le nombre d'individus risqués au temps t_i est égal au nombre d'individus risqués en t_{i-1} diminué de l'effectif des individus ayant connu l'évènement en t_{i-1} et de celui des individus censurés entre t_{i-1} inclus et t_i exclu.

Une fois ces tables construites pour l'ensemble des temps d'évènement, ensemble obtenu par l'union des deux sous-ensembles de temps d'évènement afférents à chacun des groupes (les temps considérés sont donc soient observés dans le groupe 1 soient observés dans le groupe 2), on calcule 4 quantités :

— Le nombre total d'évènements observés dans le premier groupe, c.a.d. la somme des d_{1i}, noté O_1.
— Le nombre total d'évènements observés dans le second groupe, c.a.d. la somme des d_{2i}, noté O_2.
— Le nombre total d'évènements espérés sous H0 pour le premier groupe, c.a.d. la somme des e_{1i}, noté E_1.
— Le nombre total d'évènements espérés sous H0 pour le second groupe, c.a.d. la somme des e_{2i}, noté E_2.

On note au passage la relation d'égalité $O_1 + O_2 = E_1 + E_2$.

La statistique $O_1 - E_1$ est statistique de log-rank ou de Mantel-Haenzel. Si on imagine qu'une action marketing a été effectuée auprès des individus du groupe 2 et que l'évènement est la rupture de la relation client, alors une valeur positive de la statistique signifie que le nombre d'événements observés dans le groupe 1 est supérieur à celui attendu sous H0 où, de façon équivalente puisque $O_1 - E_1 = E_2 - O_2$, que le nombre d'événements observés dans le groupe 2 est inférieur à celui attendu sous H0 et que donc l'action marketing en question affecte globalement favorablement la courbe de survie et donc la fidélisation du client. Nous reviendrons ultérieurement sur la signification du terme *globalement*.

Il reste à dériver la distribution de cette statistique pour être en mesure de réaliser un test de significativité de l'intervention. Pour cela on retrouve une statistique de Chi2 de Pearson couramment employée dans l'étude des tableaux de contingence (Cf. cours de statistique non paramétrique) :

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = (O_1 - E_1)^2 \left[\frac{1}{E_1} + \frac{1}{E_2}\right]$$

Sous l'hypothèse nulle d'égalité des courbes de survie, cette quantité est asymptotiquement la réalisation d'un Chi2 à un degré de liberté.

Dans la proc LIFETEST, SAS évalue autrement la variance de la statistique de Mantel-Haenzel : sous l'hypothèse nulle, chacun des termes $d_{1i} - e_{1i}$ est centré sur zéro et a pour variance $v_i = [n_{1i} n_{2i} d_i (n_i - d_i)]/[n_i^2 (n_i - 1)]$. Ce résultat provient du fait que, conditionnellement à d_i, d_{1i} a une distribution hypergéométrique [12]. On peut encore montrer que la variance de la somme $O_1 -$

12. Cette distribution est relative au nombre de succès dans une succession de t tirages sans remplacement. Pour mémoire, la distribution binomiale considère une succession de tirages avec remplacement. Une variable hypergéométrique

$E_1 = \sum(d_{1i} - e_{1i})$ est approximativement égale à la somme des variances de chacun des termes la constituant et que $O_1 - E_1$ tend vers une gaussienne. Dans ces conditions sous H0 :

$$\frac{O_1 - E_1}{(\sum \nu_i)^{1/2}} \rightarrow N(0,1)$$

ou encore,

$$\frac{(O_1 - E_1)^2}{\sum \nu_i} \rightarrow \chi(1).$$

, *Exemple 3* : Les données suivantes, reprises de Freireich et alii. (1963), décrivent les temps de survie (employé ici au sens littéral) de patients leucémiques avec traitement 6-MP (groupe 1, 21 patients) et sans traitement (groupe 2, 21 patients). Le signe * signale une donnée censurée.

— Groupe 1 : 6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
— Groupe 2 : 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22,23
On obtient alors :

t_i	d_{1i} (a)	n_{1i} (b)	d_{2i} (c)	n_{2i} (d)	d_i (e)=(a)+(c)	n_i (f)=(b)+(d)	e_{1i} =(b)x(e)/(f)	e_{2i} =(d)x(e)/(f)
1	0	21	2	21	2	42	1	1
2	0	21	2	19	2	40	1.05	0.95
3	0	21	1	17	1	38	0.553	0.447
4	0	21	2	16	2	38	1.135	0.865
5	0	21	2	14	2	37	1.2	0.8
6	3	21	0	12	3	35	1.909	1.091
7	1	17	0	12	1	29	0.586	0.414
8	0	16	4	12	4	28	2.286	1.714
10	1	15	0	8	1	23	0.652	0.348
11	0	13	2	8	2	21	1.238	0.762
12	0	12	2	6	2	18	1.333	0.667
13	1	12	0	4	1	18	0.75	0.25
15	0	11	1	4	1	15	0.733	0.267
16	1	11	0	3	1	14	0.786	0.214
17	0	10	1	3	1	13	0.769	0.231
22	1	7	1	2	2	9	1.556	0.444
23	1	6	1	1	2	7	1.714	0.286
$O_1 =$	9	$O_2 =$	21				$E_1 = 19.25$	$E_2 = 10.75$

X de paramètres n,s,t où n est le nombre total d'évènements et s le nombre de succès parmi ces n, vérifie

$$Prob(X = k) = \frac{\binom{s}{k}\binom{n-s}{t-k}}{\binom{n}{t}}, \text{avec} \binom{a}{b} = \frac{a!}{b!(a-b)!}.$$

Son espérance est donnée par $\frac{ts}{n}$ et sa variance par $\frac{ts(n-t)(n-s)}{n^2(n-1)}$. Dans ce qui nous intéresse ici, on considère que l'on a, à chaque date d'évènement, n_{1i} tirages, soit $t = n_{1i}$, que le nombre de succès est d_i parmi un total de n_i éléments, et comme par construction $n_{2i} = n_i - n_{1i}$, on obtient la formule donnée dans le texte.

Valant $O_1 - E_1 = -10.25$, la valeur négative de la statistique de LogRank signale que le traitement affecte favorablement le temps de survie des patients traités. Par ailleurs, le Chi2 de Pearson associé est :

$$\frac{(9 - 19.25)^2}{19.25} + \frac{(21 - 10.75)^2}{10.75} = 15.23$$

et si on compare ce chiffre aux valeurs critiques afférentes aux seuils de risque usuels de la distribution de Chi2 à un degré de liberté, on conclut que l'avantage du traitement est significatif [13].

La comparaison de k fonctions de survie, $k \geq 2$

L'extension à k groupes de la version approchée par le chi2 de Pearson du test de Mantel-Haenzel test est immédiate (ici encore, voir le cours de statistique non paramétrique). Sous l'hypothèse nulle d'égalité des k courbes de survie la quantité

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

est distribuée selon un chi2 à $k - 1$ degrés de liberté.

L'extension à k groupes du chi2 de Mantel-Haenzel est également possible. Pour cela il est nécessaire de calculer la matrice de variance-covariance de seulement $k - 1$ termes arbitrairement pris parmi les k statistiques $O_1 - E_1$, $O_2 - E_2$, $\ldots$, $O_k - E_k$. Sous l'hypothèse nulle, un théorème standard assure que la forme quadratique construite à partir de ce vecteur sur l'inverse de leur matrice de variance covariance est la réalisation d'un Chi2 à $k - 1$ degrés de liberté [14].

En cas de rejet de l'hypothèse nulle, on peut être amené à rechercher les couples de fonctions responsables de ce rejet. L'hypothèse nulle rejetée étant une hypothèse jointe de la forme $H0$: $S_1(t) = S_2(t) = \cdots = S_k(t)$, on va retrouver la difficulté habituelle de contrôle du risque global et donc du nécessaire ajustement du seuil de risque à utiliser pour chacun des tests individuels. Comme on le sait, la méthode la plus courante (et la plus simple) est celle de Bonferroni : pour un seuil de risque α fixé a priori, si n_H hypothèses simples doivent être considérées, alors on rejettera l'hypothèse nulle $H0 : S_i(t) = S_j(t)$ si son seuil de significativité est inférieur à α/n_H ou, de manière équivalente, si son seuil de significativité multiplié par n_H est inférieur à α [15]. D'autres méthodes d'ajustement sont également disponibles. Ainsi l'ajustement de Sidák affiche $1 - [1 - SL_b]^{n_H}$ que l'on compare toujours à α pour prendre la décision de rejet ou non. Il s'agit de la méthode par défaut utilisée par proc LIFETEST [16].
Partant de l'hypothèse jointe précédente, le nombre d'hypothèses simples, n_H, est constitué de l'ensemble des couples pouvant être constitués, soit $n_H = \frac{k(k-1)}{2}$. Il est cependant possible de se donner une fonction de survie de référence pour ne considérer que les écarts à cette référence

13. Si on emploie l'autre mode de calcul alors la variance de la statistique de Mantel-Haenzel est estimée à 6.257, et le Chi2 associé est $10.251^2/6.257 = 16.79$ ce qui mène ici à la même conclusion que le chi2 de Pearson

14. Pour les formules explicites de construction de cette matrice, voir la documentation de la proc LIFEREG.

15. dans la proc LIFETEST, c'est cette dernière convention qui est utilisée : si on note SL_b le seuil de significativité brut d'un test simple, l'affichage des résultats fait apparaître SL_b lui-même et le seuil ajusté égal à $\min(1, n_H \times SL_b)$.

16. Voir la documentation de cette procédure pour les autres choix possibles.

et dans ce cas $n_H = k - 1$. Les deux possibilités [17] sont offertes dans la proc LIFETEST avec des variantes (plusieurs courbes de référence, liste de couples à comparer,..).

2.8.2 Le test de Wilcoxon (ou de Gehan) et les autres statistiques pondérées

Le test du log-rank a donc pour expression (pour 2 groupes) $\sum_{i=1}^{r}(d_{1t_i} - e_{1t_i})$ où r est le nombre d'évènements observés sur les groupes 1 et 2. Implicitement il attribue un poids unitaire à chacune des quantités $d_{1t_i} - e_{1t_i}$. On peut imaginer de construire des statistiques pondérées de la forme :

$$\sum_{i=1}^{r} w_i(d_{1t_i} - e_{1t_i})$$

Ceci permet en jouant sur la valeur des coefficients de pondération w_i d'attribuer plus ou moins d'influence aux évènements en fonction de la durée de leurs réalisations. Ainsi une décroissance de ces poids accorde plus d'influence aux évènements de courte durée. Une proposition a été faite par Gehan (1965) à l'origine pour deux groupes, qui est en fait une généralisation du test de Wilcoxon, et a été étendue à k groupes par Breslow (1970). On pose simplement $w_i = n_i$, c'est-à-dire que les poids sont égaux au nombre d'individus risqués au temps t_i. Comme n_i diminue avec t_i, cette statistique accorde donc plus de poids aux évènements de courte durée relativement au test de Mentel-Haenzel [18].

Dans la proc LIFETEST, le test Wilcoxon-Gehan est fourni par défaut au même titre que le test de logrank. On peut également mettre en oeuvre d'autres tests qui se fondent sur des coefficients de pondération w_i différents, ainsi Tarone-Ware avec $w_i = \sqrt{n_i}$ attribue un poids intermédiaire entre celui du logrank et celui de Wilcoxon-Gehan. Peto-Peto utilise l'estimation de la survie, soit $w_i = \hat{S}(t_i)$. Celle-ci étant non croissante, cela revient à attribuer plus de poids aux écarts de survie observés aux temps d'évènement faibles. Cette sur-représentation des temps les plus courts est encore accentuée avec le 'modified Peto-Peto'. Enfin, la version de Harrington-Fleming(p,q) en posant $w_i = \hat{S}(t_i)^p(1 - \hat{S}(t_i))^q$, $p, q \geq 0$ permet de se concentrer sur certains sous-espaces du support de la survie [19]

Remarques

— Sous certaines conditions et notamment si le ratio des taux de risque est constant alors le log-rank a le plus fort pouvoir dans la classe des tests de rangs linéaires (Peto et Peto, 1972). Sous cette hypothèse de risque proportionnel, que nous détaillerons par la suite (Chapitre 3), les fonctions de survie $S_1(t)$ et $S_2(t)$ des individus appartenant à deux classes différentes

17. Par exemple, avec k=3, en cas de rejet de $H0 : S_1(t) = S_2(t) = S_3(t)$, alors dans le premier cas on est amené à regarder $H0 : S_1(t) = S_2(t)$, $H0 : S_1(t) = S_3(t)$, $H0 : S_2(t) = S_3(t)$ et donc $n_H = 3$. Dans le second, si $S_1()$ est prise comme référence, on aura seulement $H0 : S_1(t) = S_2(t)$ et $H0 : S_1(t) = S_3(t)$ avec $n_H = 2$.

18. Naturellement il faut adapter les expressions des variances ou des variances-covariances pour tenir compte de la présence des poids w_i.

19. Ainsi par exemple, avec p proche de 1 et q proche de zéro on se concentre sur les écarts existant aux temps d'évènements faibles, retrouvant ainsi à la limite les pondérations de Peto-Peto. Avec p proche de zéro et q proche de 1 on va accorder plus de poids aux écarts existants aux temps d'évènements élevés (alors que la survie est la plus faible). Avec $p = 1/2$ et $q = 1/2$, ce sont les écarts observés pour des survies aux environ de 0.5 qui sont sur-pondérés.

i et j vérifient $S_1(t) = S_2(t)^k$. En conséquence on obtient des courbes parallèles dans l'espace $\log - \log(S_t)$ versus log(t). Un simple graphique peut donc, visuellement, permettre de voir si l'hypothèse en question est raisonnable ou pas [20].

— Les tests les plus couramment utilisés sont ceux du logrank et de Wilcoxon-Gehan. A moins d'avoir de bonnes raisons de faire autrement, on conseille généralement de les considérer en priorité. Dans tous les cas il est important de ne pas fonder le choix du test ex-post à la vue des résultats : les conclusions que l'on est amené à tirer selon les différents tests peuvent se contredire et il serait alors possible de valider n'importe quelle conjecture. Il est donc important de faire ce choix ex-ante compte-tenu notamment de la perception a priori que l'on a de la validité de l'hypothèse de risques proportionnels ou des plages sur lesquelles la divergence des survies est la plus intéressante à considérer.
— Ces tests requièrent que la distribution des censures ne soit pas trop déséquilibrée entre les différentes sous-populations.
— Lorsque les courbes de survie se coupent alors la puissance des tests peut être affectée, ceci évidemment en raison d'un effet de compensation algébrique qui se produit dans le calcul de la somme des quantités $d_{1t_i} - e_{1t_i}$. Par ailleurs l'intersection des courbes remet en cause l'hypothèse de risque proportionnel [21] et donc l'optimalité du test de log-rank.
— Lorsque les effectifs des individus à risque diminuent, la précision des estimateurs se dégrade. Il est donc recommandé de surveiller l'évolution de ces effectifs avec l'augmentation des temps de survie afin de s'assurer qu'un nombre raisonnable d'observations sont utilisées dans la construction des estimateurs de la survie et des tests de comparaison.

Un premier exemple

Pour illustrer les développements qui précédent, nous prenons les données de Lee(1992) : il s'agit de comparer l'efficacité de deux traitements d'immunothérapies (BCG vs. Cryptosporium parvum) sur la survie de patients développant un mélanome malin. Pour chaque patient on connaît la nature du traitement, le temps de survie censuré ou non (une étoile signale un temps censuré) ainsi que son appartenance à une classe d'âge. Ces informations sont présentées dans le tableau (2.2).

Via une étape data on a créé la variable treat valant 0 si le patient a reçu le traitement BCG et 1 sinon, ainsi que la variable c valant 1 si le patient est décédé, 0 sinon. Enfin, la variable time contient les différentes durées. Le programme suivant qui autorise la récupération de graphiques en format postscript est ensuite exécuté :

```
proc lifetest plots=(s(nocensor atrisk) lls);
time time*c(0);
strata treat / test=(logrank wilcoxon);
run;
```

20. Admettons la proportionnalité des risques : $h_1(t) = kh_2(t)$ où k est une constante positive. En utilisant les relations fondamentales, il vient :$\log S_1(t) = -H_1(t) = -\int_0^t h_1(u)du = -\int_0^t kh_2(u)du = -kH_2(t) = k\log S_2(t)$, soit encore $S_1(t) = S_2(t)^k$. En conséquence, $\log - \log S_1(t) = \log k + \log - \log S_2(t)$ assurant ainsi le parallélisme des courbes dans l'espace $\log - \log S(t)$ versus t ou encore versus $log(t)$.

21. Cela découle directement des éléments présentés dans la note 20.

21-40		41-60		61+	
BCG	C. parvum	BCG	C. parvum	BCG	C. parvum
19	27*	34*	8	10	25*
24*	21*	4	11*	5	8
8	18*	17*	23*		11*
17*	16*		12*		
17*	7		15*		
34*	12*		8*		
	24		8*		
	8				
	8*				

TABLE 2.2 – Efficacité de deux traitements - Lee (1992)

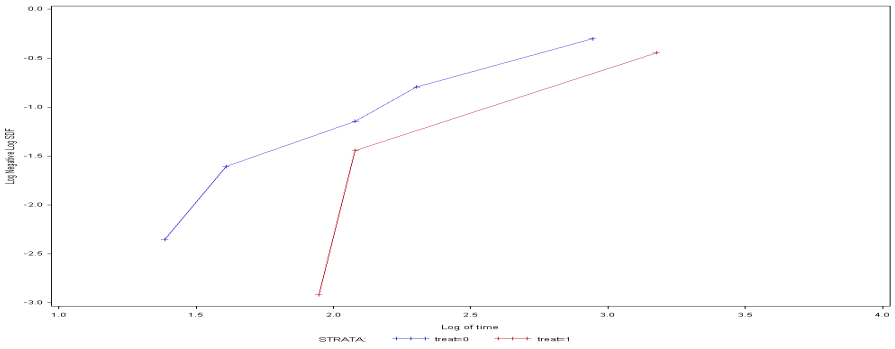


FIGURE 2.3 – Log-log S(t) - Lee (1992)

Les distributions des censures sont assez différentes dans les deux échantillons de patients. Par ailleurs, le graphique (2.3) des deux transformées des survies estimées montre que le parallélisme n'est pas vérifié sur les temps les plus faibles, ce qui peut affecter le test de logrank. Il peut donc être utile de le compléter par un autre test, ici Wilcoxon-Gehan. Les deux courbes de survie obtenues sont par ailleurs présentées dans le graphique (2.4).

Les résultats des tests demandés sont présentés dans le tableau 2.3 et conduisent à ne pas rejeter, aux seuils de risque usuels, l'hypothèse d'égalité des survies et donc l'équivalence en termes d'efficacité des deux traitements.

	Statistique	Chi2	df	SL
Log-rank	1.2893	0.7558	1	0.3847
Wilcoxon	34.000	0.9115	1	0.3397

TABLE 2.3 – Tests d'égalité des survies - données Lee (1992)

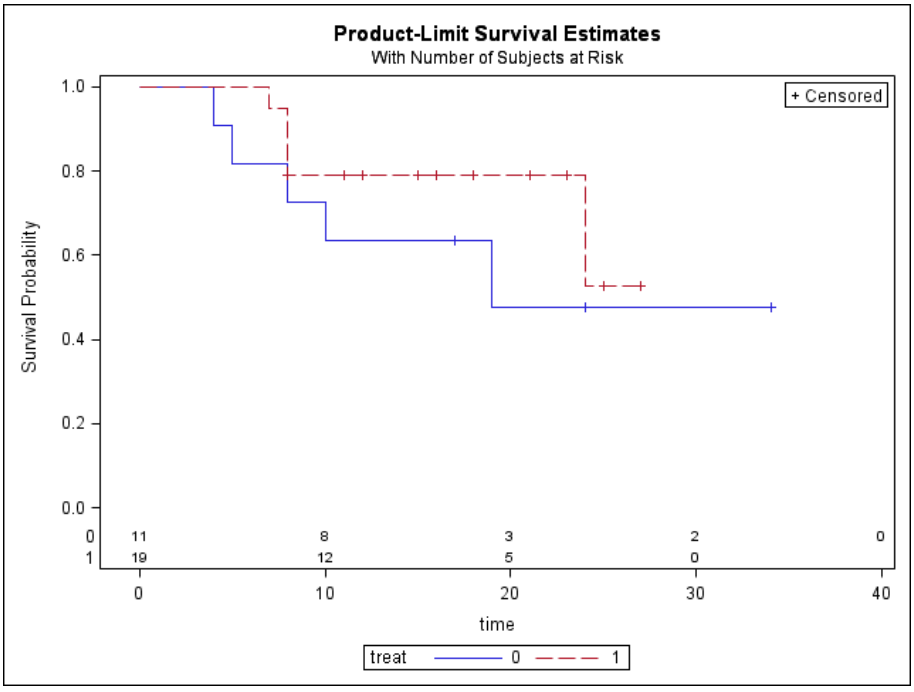


FIGURE 2.4 – Courbes de survie estimées - Lee (1992)

Un deuxième exemple

On reprend les données du fichier BMT, les patients atteints de leucémie sont distingués selon la nature de celle-ci : ALL (acute lymphoblastic leukemia), AML (acute myeloctic leukemia)-Low Risk, et AML-High Risk. La question est de comparer les survies entre ces trois groupes. L'exécution du code suivant réclame la construction des tests par défaut (logrank et de Wilcoxon) d'égalité des 3 courbes ainsi que le calcul des tests d'égalité pour tous les couples possibles, ici 3, avec un ajustement de Bonferroni sur le seuil de significativité. Le graphique 2.5 contient la représentation des courbes de survie estimées au sein de chaque groupe.

```
proc lifetest data=BMT plots=survival(atrisk=0 to 2500 by 500);
time T * Status(0);
strata Group / adjust=bon;
run;
```

Les résultats des tests d'égalité des trois courbes et des tests simples sont respectivement dans les tableaux 2.4 et 2.5. Avec un seuil de risque de 10% on serait conduit à rejeter l'hypothèse d'égalité jointe. Les tests d'hypothèses simples quand à eux permettent d'accepter l'homogénéité des survies des patients appartenant aux groupes ALL et AML-High Risk et à les distinguer de celles afférentes aux patients classés dans le groupe des ALM-Low risk, ces derniers étant favorisés au regard du temps de survenue du décès.

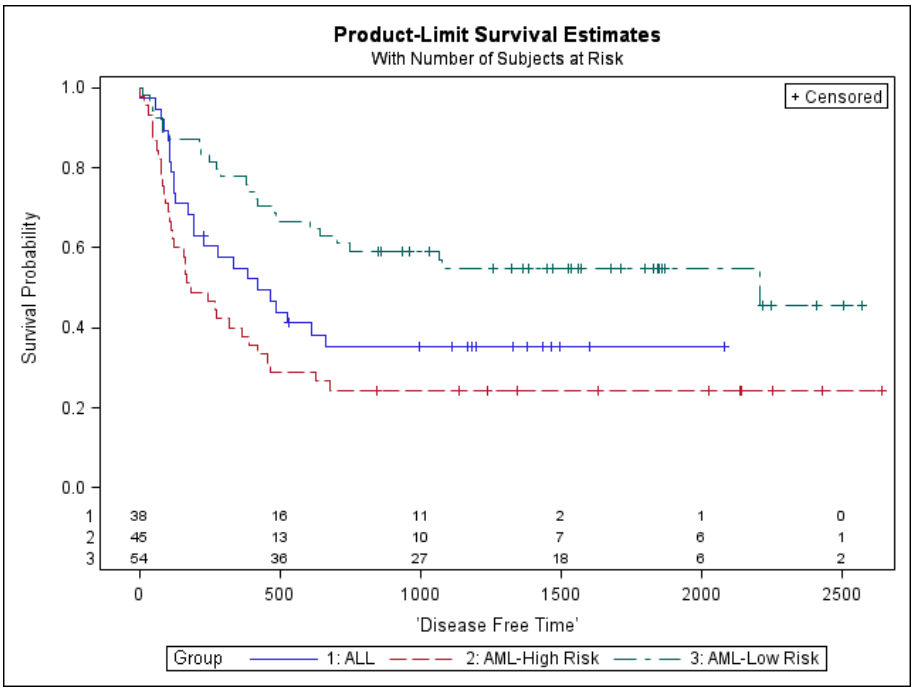


FIGURE 2.5 – Courbes de survie estimées - Données BMT

Test	Chi2	df	SL
Log-rank	13.8037	2	0.0010
Wilcoxon	16.2407	2	0.0003

TABLE 2.4 – Tests d'égalité des 3 survies - données BMT

		Test de Log-rank		
Groupe	Groupe	Chi2	SL brut	SL Bon.
ALL	AML-High Risk	2.6610	0.1028	0.3085
ALL	AML-Low Risk	5.1400	0.0234	0.0701
AML-High Risk	AML-Low Risk	13.8011	0.0002	0.0006
		Test de Wilcoxon		
Groupe	Groupe	Chi2	SL brut	SL Bon.
ALL	AML-High Risk	3.8056	0.0511	0.1532
ALL	AML-Low Risk	5.1415	0.0234	0.0701
AML-High Risk	AML-Low Risk	16.2052	<0.0001	0.0002

TABLE 2.5 – Tests d'égalité des survies entre tous les couples - données BMT

2.8.3 Les tests stratifiés de comparaison des survies

Il peut arriver que l'on soupçonne une hétérogénéité des sous-populations constituant chacune des strates que l'on veut comparer. Si cette hétérogénéité est bien présente, alors les tests précédents peuvent être déficients et leurs conclusions sujettes à caution. Supposons par exemple que l'on veuille comparer la durée d'accès à l'emploi de deux filières de formation, mais que la répartition par sexe des sortants de ces deux filières soit différentes. Dans ces conditions, nous pourrions attribuer aux filières ce qui en fait relève de conditions d'accès à l'emploi éventuellement inégales entre les hommes et les femmes.
Les tests stratifiés visent à tenir compte de ce type d'hétérogénéité. Si on note X1 la variable dont les modalités définissent les strates d'intérêt et X2 celle dont les modalités définissent les sous-populations éventuellement hétérogènes au sein des strates précédentes, alors la logique des tests stratifiés est de construire des tests d'égalité des survies sur chacune des sous-populations responsables de l'hétérogénéité, identifiées par X2, puis de combiner les valeurs de ces tests pour donner un test global d'égalité des survies des strates identifiées par X1.
Dans l'exemple précédent, on commencerait ainsi par tester l'égalité des courbes de survie des femmes entre les deux filières de formation F1 et F2, puis l'égalité des survies des hommes selon leur appartenance à F1 ou F2 pour construire une statistique globale d'égalité des survie entre F1 et F2 à partir des valeurs obtenues aux deux tests précédents.

Plus généralement, si on a M strates de sous-populations pouvant créer de l'hétérogénéité on récupère alors à la première étape M statistiques v_s, chacune de variance-covariance estimée V_s et de degrés de liberté df. A la seconde étape, on construit les quantités

$$\begin{aligned} v &= \sum_{s=1}^{M} v_s \text{ et} \\ V &= \sum_{s=1}^{M} V_s \end{aligned}$$

Finalement, la statistique de test stratifiée est construite comme $v' V^{-1} v$ et, sous l'hypothèse d'égalité des survie, elle possède asymptotiquement une distribution de Chi2 à df degrés de liberté.

Un exemple

On reprend les données de Lee (1992) déjà utilisées. Il s'agissait d'étudier l'efficacité comparée de deux traitements (BCG vs. Cryptosporidium parvum). Dans cette base nous avons également la répartition des patients en 3 classes d'âge, information qui n'avait pas été utilisée précédemment. On peut imaginer que l'efficacité d'un traitement soit affectée par l'âge du patient. Si tel est le cas, alors on pourrait attribuer à l'un des traitement ce qui ne serait qu'une conséquence de répartition par âge hétérogène entre les deux échantillons, ou bien au contraire masquer la plus grande efficacité de l'un des traitement en raison d'une répartition par âge déséquilibrée. Dans tous les cas, le risque d'avoir une mauvaise appréciation de leur efficacité relative peut être élevé.

Test	Statistique	Chi2	df	SL
Wilcoxon	6	0.1786	1	0.6726

TABLE 2.6 – Test stratifié d'égalité des survies - Lee (1992)

	agegrp=1		agegrp=2		agegrp=3	
Test	Stat.	Variance	Stat.	Variance	Stat.	Variance
Wilcoxon	-3.0	155.615	5.0	35.000	4.0	11.000

TABLE 2.7 – Test stratifié d'égalité des survies - Lee (1992)

L'exécution des lignes suivantes demande le calcul de la statistique de Wilcoxon stratifiée correspondante et conduisent aux résultats présentés dans le tableau 2.6.

```
proc lifetest;
time time*c(0);
strata agegrp /test=wilcoxon group=treat;
run;
```

Pour bien comprendre la construction de cette statistique, on peut détailler les diverses étapes. En premier lieu, des statistiques d'égalité des survies de patients soumis à des traitements différents mais appartenant à une même classe d'âge sont réclamées par les instruction suivantes :

```
proc sort;
by agegrp;
run;
proc lifetest;
by agegrp;
time time*c(0);
strata treat /test=wilcoxon;
run;
```

Ces statistiques ont ici un seul degré de liberté et les résultats sont donnés dans le tableau 2.7. Finalement, on construit la statistique de test stratifiée comme indiquée plus haut pour obtenir, avec $v = (-3.0+5.0+4.0) = 6.0$ et $V = (155.615+35.0+11.0) = 201.615$, une valeur de $\frac{6^2}{201.615} = 0.1786$ qui est, sous l'hypothèse nulle, la réalisation d'un Chi2 à 1 degré de liberté comme indiqué dans le tableau 2.6 [22].

22. Cette façon de procéder réconcilie les résultats apparemment contradictoires de SAS et Stata sur les tests stratifiés et qui avaient été relevés dans une note de la FAQ de Stata *'Why do Stata and SAS differ in the results that they report for the stratified generalized Wilcoxon test for time-to-event data ?'* disponible ici : http ://www.stata.com/support/faqs/stat/wilcoxon.html. La combinaison de la commande STRATA avec l'option GROUP donne bien des résultats identiques à ceux obtenus avec la commande 'sts test treat, wilcoxon strata(agegrp)' de Stata. L'origine de la contradiction était l'emploi sous SAS, par les auteurs de la note en question, de la commande STRATA couplée avec la commande TEST. A leur décharge, il est vrai que l'option GROUP n'existait pas dans la version 8 de SAS.

2.8.4 Tests d'association entre une variable continue et la survie

Les tests précédents permettent de juger de l'influence sur la survie de variables de type nominal, ordinal ou numérique ayant relativement peu de modalités. Dans le cas de variables continues, le nombre de strates que l'on peut être amené à construire est susceptible de les rendre inapplicables. On propose alors de construire des tests de rang de la façon suivante : si m variables numériques $\mathbf{z} = (z_1, z_2, \ldots, z_m)'$ sont considérées, la statistique de test est construite comme

$$v = \sum_{i=1}^{n} s_{(i,c_i)} \mathbf{z}_i$$

où n est le nombre total d'observations et s le score associé à la $i^{\text{ème}}$ observation qui dépend du mécanisme de censure via c_i. Dans SAS 9.2, la proc LIFETEST considère soit des scores de log-rank, soit des scores de Wilcoxon. Une matrice de variance-covariance V est évaluée [23] et finalement deux types d'information sont fournis :

— Le premier concerne des statistiques individuelles : pour chacune des m variables une statistique est construite comme $v(i)^2/V_{ii}$ et comparée à la valeur critique tirée d'une distribution de Chi2 à 1 degré de liberté. Le rejet de l'hypothèse nulle laisse penser que la survie dépend de la variable considérée. Cette procédure est souvent employée pour identifier les variables pertinentes à retenir pour une explication de la survie ou du risque dans les modèles paramétriques (Chapitre 3) ou dans le modèle de Cox (Chapitre 4).

— La seconde information est relative aux résultats d'une procédure de sélection de type Forward. On calcule la statistique usuelle de test global $v'V^{-1}v$ en s'arrangeant dans la méthode du pivot pour faire apparaître l'ordre des contributions de chaque variable à cette statistique. Si on note $z_{(i)}$ les variables classées selon cette procédure, alors $z_{(1)}$ est celle ayant la plus grande contribution puis $z_{(2)}$ celle qui, associée à $z_{(1)}$ contribue le plus à l'augmentation de la statistique, etc.... Une autre façon de comprendre cette démarche est de concevoir que si on voulait expliquer la survie par un modèle linéaire, alors parmi les m modèles ayant une seule explicative (z_1 ou z_2 ou ... z_m), le R^2 est maximal si on prend $z_{(1)}$. Parmi les modèles à deux explicatives, celui ayant le plus grand R^2 est constitué de $z_{(1)}$ et $z_{(2)}$, etc.... Là encore l'objectif est souvent d'aider à la sélection d'une liste de variables continues pertinentes avant l'ajustement d'un modèle paramétrique ou semi-paramétrique.

Comme vu dans la section précédente, et pour des raisons identiques, ces tests d'associations peuvent également être stratifiés. Dans ce cas, une statistique similaire à celle que l'on vient d'exposer est tout d'abord calculée sur chacune des strates considérées puis, dans la deuxième étape, ces statistiques sont combinées pour construire le test final.

En pratique ces statistiques de rang nécessitent que les nombre d'évènements simultanés et de censures ne soient pas trop importants relativement au nombre total d'observation.

23. Voir la documentation de LIFETEST pour le détail des expressions

	Logrank			Wilcoxon		
Variable	Stat.	Chi2	SL	Stat.	Chi2	SL
Age	-83.7764	0.6107	0.4345	-12.6165	0.0320	0.8581
Prior	36.2253	0.4980	0.4804	-5.7669	0.0389	0.8436
DiagTime	-93.9987	1.0011	0.3170	-65.2381	0.8031	0.3702
Kps	1220.1	44.8525	<.0001	920.4	57.4490	<.0001
Treatment	-0.5002	0.00817	0.9280	-3.1226	0.8782	0.3487

TABLE 2.8 – Chi2 univariés, statistiques de Logrank et de Wilcoxon

Un exemple sans stratification

Afin d'illustrer les développements qui précèdent, nous reprenons un fichier de données disponible dans l'aide de la proc LIFETEST [24]. La variable SurvTime contient les temps de survie en jours de patients ayant un cancer de la gorge. Lex explicatives sont Cell (type de tumeur), Therapy (type de thérapie : standard ou test), Prior (existence d'une thérapie antérieure : 0=non, 10=oui), Age (âge en années), DiagTime (temps en mois de la date du diagnostic à l'entrée dans l'échantillon), Kps (indicateur de performance mesuré par l'indice de karnofsky). La variable de censure, Censor, est créée et vaut 1 si la durée est censurée et 0 si le décès est observé. La variable indicatrice du traitement, Treatment, prend la valeur 0 pour la thérapie standard et la valeur 1 pour la thérapie en test.

Dans ce qui suit, on teste l'impact des variables Age, Prior, DiagTime, Kps et Treatment sur la survie.

```
proc lifetest data=VALung;
time SurvTime*Censor(1);
test Age Prior DiagTime Kps Treatment;
run;
```

On récupère alors les résultats donnés dans le tableau 2.8 pour ce qui concerne les statistiques individuelles et dans le tableau 2.9 pour la procédure de sélection. Notez que dans le premier l'ordre de présentation est celui qui est donné par la commande 'test' du programme d'appel, dans le second les variables sont classées en fonction de leur contribution à la construction de la statistique globale $v'V^{-1}v$. On rappelle également que le nombre de degré de liberté du Chi2 est toujours de 1 dans le premier tableau alors qu'il est égal au nombre de variables incorporées dans le meilleur modèle dans le second. Ainsi, la variable la plus pertinente pour séparer les survies est KPS. Si on veut retenir deux variables, celles-ci seraient Kps et DiagTime, etc…. Notez aussi que le nombre de degré de liberté s'accroît d'une unité avec l'ajout d'une variable supplémentaire indépendamment du nombre de valeurs différentes que prend la variable en question. Enfin, vous remarquerez également que cet exemple illustre la non équivalence des tests Logrank et Wilcoxon.

Un exemple avec stratification

Pour illustrer la construction d'une statistique stratifiée, on reprend les données de l'exemple précédent. L'idée a priori est que l'influence de nos cinq variables peut être différente selon la nature de la tumeur (variable Cell discriminant les tumeurs entre 4 types : squamous, small, adeno

24. Example 49.1 : Product-Limit Estimates and Tests of Association.

	Logrank				Wilcoxon			
DF	Variable	Chi2	SL	△Chi2	Variable	Chi2	SL	△Chi2
1	Kps	44.8525	<.0001	44.8525	Kps	57.4490	<.0001	57.4490
2	Treatment	46.2596	<.0001	1.4071	Age	58.5609	<.0001	1.1119
3	Prior	46.6821	<.0001	0.4225	DiagTime	58.7809	<.0001	0.2200
4	DiagTime	46.7795	<.0001	0.0974	Treatment	58.8881	<.0001	0.1072
5	Age	46.8386	<.0001	0.0591	Prior	58.9000	<.0001	0.0120

TABLE 2.9 – Procédure de sélection Stepwise, statistiques de Logrank et de Wilcoxon

	Logrank			Wilcoxon		
Variable	Stat.	Chi2	SL	Stat.	Chi2	SL
Age	-40.7383	0.1485	0.7000	14.4158	0.0466	0.8290
Prior	-19.9435	0.1802	0.6712	-26.3997	0.8336	0.3612
DiagTime	-115.9	1.4013	0.2365	-82.5069	1.3127	0.2519
Kps	1123.1	43.4747	<.0001	856.0	51.9159	<.0001
Treatment	-4.2076	0.6967	0.4039	-3.1952	1.0027	0.3167

TABLE 2.10 – Chi2 univariés, statistiques de Logrank et de Wilcoxon avec stratification

et large).

Le programme à exécuter devient le suivant, avec comme résultats les chiffres des tableaux 2.10 et 2.11. Dans le présent cas, la prise en compte ou non d'une stratification ne modifie pratiquement pas les conclusions : parmi les cinq variables considérées, seule Kps paraît devoir être retenue comme explicative de la survie, les accroissements de la statistique générés par les autres variables pouvant être considérés comme négligeables [25]

```
proc lifetest data=VALung;
time SurvTime*Censor(1);
strata Cell;
test Age Prior DiagTime Kps Treatment;
run;
```

	Logrank				Wilcoxon			
DF	Variable	Chi2	SL	△Chi2	Variable	Chi2	SL	△Chi2
1	Kps	43.4747	<.0001	43.4747	Kps	51.9159	<.0001	51.9159
2	Treatment	45.2008	<.0001	1.7261	Age	53.5489	<.0001	1.6329
3	Age	46.3012	<.0001	1.1004	Treatment	54.0758	<.0001	0.5269
4	Prior	46.4134	<.0001	0.1122	Prior	54.2139	<.0001	0.1381
5	DiagTime	46.4200	<.0001	0.00665	DiagTime	54.4814	<.0001	0.2674

TABLE 2.11 – Procédure de sélection Stepwise, statistiques de Logrank et de Wilcoxon avec stratification

25. Même si ceci n'est pas totalement justifié, puisque l'on est dans une procédure de sélection pas à pas, une règle simple consiste à comparer ces accroissements à la valeur critique d'un Chi2 à 1 degré de liberté. Que l'on considère ou non une stratification selon la nature de la tumeur, on constate qu'aucun des accroissements n'est significatif. Ceci rejoint parfaitement les conclusions que l'on pouvait tirer de l'examen des statistiques univariées.

mois	janvier	février	mars	avril	mai	juin	non revenus
	1	2	3	4	5	6	fin juin
réachats	12	45	58	38	19	10	88

TABLE 2.12 – Répartition des réapprovisionnements sur les 6 mois suivants le premier achat

2.9 Un exemple de fonction de risque en temps discret : étude du comportement de réapprovisionnement

Le cas simple que nous allons traiter s'inspire d'un exemple proposé par Allison. Il doit aider à comprendre l'intérêt de l'estimation de la fonction de risque. La question posée est celle d'une meilleure connaissance du comportement de réapprovisionnement d'un produit quelconque par les clients d'une enseigne commerciale afin de conseiller des actions marketing visant par exemple à favoriser le retour du client ou encore à le fidéliser.
On suppose qu'en un mois donné, par exemple en décembre, 270 clients ont acheté un certain produit. Parmi ceux-ci, certains ont renouvelé leur achat en janvier, d'autres en février, etc. La table 2.12 présente la répartition de ces renouvellements sur les 6 mois suivant la première acquisition.

Le "risque" de se réapprovisionner au cours d'un mois donné est donc la probabilité d'effectuer un nouvel achat au cours de ce mois sachant qu'on ne l'a pas fait auparavant. L'évaluation de ce risque est donc naturellement donnée par :

$$h(mois) = \frac{\text{effectif ayant renouvelé au cours du mois}}{\text{effectif qui aurait été en mesure de renouveler au cours du mois}}$$

Ainsi, dans cet exemple et en temps discret, on obtient les valeurs présentées dans la table 2.13. La procédure `Lifetest` ne permet pas le calcul direct de ces risques évalués en temps discret. Il est toutefois possible de les obtenir à partir de l'estimateur de Kaplan-Meier. En effet, le raisonnement nous ayant mené à l'équation (2.1) est valable que l'on soit en temps discret ou en temps continu, et donc :

$$S(t_i) = [1 - h(t_i)] \times S(t_{i-1})$$

soit encore :

$$Pr[T \geq t_i] = (1 - Pr[T = t_i | T \geq t_i]) \times Pr[T \geq t_{i-1}]$$

ce qui signifie simplement que la probabilité de connaître l'événement en ou après t_i est égale à la probabilité de ne pas le connaître en t_i conditionnellement au fait de ne pas l'avoir connu avant multiplié par la probabilité de ne pas l'avoir connu avant t_i . Il vient donc :

$$h(t_i) = 1 - \frac{S(t_i)}{S(t_{i-1})} \tag{2.15}$$

On retrouve avec ces égalités un résultat déjà bien connu : connaissant la fonction de survie on peut retrouver la fonction de risque [26], ou bien connaissant la fonction de on peut retrouver la fonction de risque. On peut donc via `Lifereg` calculer la survie pour ensuite estimer le risque au moyen de (2.15). Pour cet exemple, on exécutera les commandes suivantes :

26. Des égalités précédentes on déduit en effet aisément que $S(t_i) = [1 - h(t_i)][1 - h(t_{i-1})] \ldots [1 - h(t_1)]$

h(janvier)	$=\frac{12}{270}$		$= 0.04444$
h(février)	$=\frac{45}{270-12}$	$=\frac{45}{258}$	$= 0.17442$
h(mars)	$=\frac{58}{258-45}$	$=\frac{58}{213}$	$= 0.27230$
h(avril)	$=\frac{38}{213-58}$	$=\frac{38}{155}$	$= 0.24516$
h(mai)	$=\frac{19}{155-38}$	$=\frac{19}{117}$	$= 0.16239$
h(juin)	$=\frac{10}{117-19}$	$=\frac{10}{98}$	$= 0.10204$

TABLE 2.13 – estimation directe du "risque" de réapprovisionnement

```
 data achats;
input mois censor eff;
cards;
1 1 12
2 1 45
3 1 58
4 1 38
5 1 19
6 1 10
6 0 88
;
run;
proc lifetest data=achats outsurv=results;
freq eff;
time mois*censor(0);
run;
data results;
set results;
h=1-survival/lag(survival);
run;
proc print data=results;
run;
proc sgplot data=results;
vbar mois / response=h;
run;
```

Les sorties des procédures `print` et `sgplot` sont présentées dans la table 2.14. Vous pouvez vérifier que les valeurs du risque de retour du client sont bien identiques à celles obtenues avec le calcul direct de la table 2.13. Le graphique facilite l'interprétation des résultats : il montre que le réapprovisionnement par les clients du produit s'effectue majoritairement 3 à 4 mois après l'achat initial. En pratique cela peut être une incitation à mener une action envers les clients concernés deux à 3 mois après après leur première acquisition afin de renforcer cette appétence. On peut aussi envisager de lancer une autre campagne 5 à 6 mois après afin de lutter contre la perte de clients qui semble se dessiner au delà de 5 mois. Pour conclure, vous pouvez encore imaginer qu'un suivi sur une plus longue période de ces comportements de réapprovisionnement puisse faire apparaître de possibles saisonnalités également exploitables par l'enseigne.

Obs	mois	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL	h
1	0	.	1.00000	1.00000	1.00000	.
2	1	0	0.95556	0.92306	0.97451	0.04444
3	2	0	0.78889	0.73520	0.83293	0.17442
4	3	0	0.57407	0.51280	0.63052	0.27230
5	4	0	0.43333	0.37366	0.49145	0.24516
6	5	0	0.36296	0.30591	0.42015	0.16239
7	6	0	0.32593	0.27078	0.38212	0.10204
8	6	1	.	.	.	.

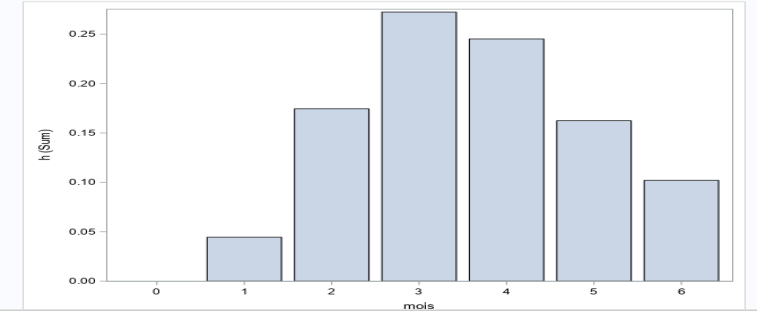


TABLE 2.14 – Risque de réapprovisionnement évalué via la survie estimée

2.10 Les tables de survie - La méthode actuarielle

La construction de tables de survie remonte au 18$^{\text{ième}}$ siècle. Elles ont été utilisées notamment par les démographes pour étudier la longévité des populations. Un autre exemple d'application concerne le domaine de l'assurance où il est important d'estimer la probabilité de décès d'un individu à différents âges pour évaluer le prix d'une police d'assurance. Elles peuvent également être utiles dans le cadre des analyses contemporaines des données de survie. C'est particulièrement le cas lorsqu'au lieu de connaître la date exacte de survenue d'un évènement ou d'une censure on ne connaît qu'un intervalle de temps dans lequel l'un ou l'autre se réalise. Un autre cas de figure est celui où le nombre de durées observées est important : alors qu'avec KM les calculs sont réalisés pour chacune de ces durées, ils ne sont effectués que pour chacun des intervalles de temps considérés dans la construction d'une table de survie.

Relativement aux méthodes de calcul employées dans l'estimateur KM, une seule différence notable intervient dans le traitement des censures au sein d'un intervalle : on va utiliser l'effectif moyen à risque pendant l'intervalle de temps considéré dans l'expression donnant la probabilité de survie. Ceci revient à supposer que la censure se produit de manière uniforme sur l'intervalle. Si l'effectif d'individus risqués est n_i au début de l'intervalle $[t_i, t_{i+1}[$ et que l'on observe sur ce laps de temps c_i données censurées alors juste avant t_{i+1} le nombre d'individus à risque est de $n_i - c_i$. L'effectif risqué moyen sur $[t_i, t_{i+1}[$ est donc [27] $n_i^* = n_i - \frac{c_i}{2}$. En conséquence la probabilité de survenue de l'évènement au cours de ce $i^{\text{ième}}$ intervalle pour les individus qui ne l'ont pas connu au cours des $(i-1)$ intervalles précédents est donnée par

$$\hat{q}_i^* = \frac{d_i}{n_i - \frac{c_i}{2}} = \frac{d_i}{n_i^*},$$

27. On fait apparaître l'exposant * pour distinguer les estimateurs construits avec les effectifs à risque corrigés n_i^* des estimateurs KM construits sur n_i.

et la probabilité de survie conditionnelle au cours du $i^{\text{ième}}$ intervalle est donc $\hat{\pi}_i^* = 1 - \hat{q}_i^*$.

A partir de là, l'estimation de la probabilité de survie s'effectue selon une récurrence faisant appel au théorème des probabilités conditionnelles et à l'évidence selon laquelle pour survivre à un intervalle donné il faut déjà avoir survécu à tous les intervalles précédents. Si on note *survie(i)* le fait de survivre, cad de ne pas connaître l'évènement, au $i^{\text{ième}}$ intervalle de temps, alors il vient :

$$\begin{aligned}
\hat{Prob}[\textit{survie(i)}] &= \hat{Prob}[\textit{survie(i)} \text{ et } \textit{survie(i-1)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&= \hat{Prob}[\textit{survie(i)} \mid \textit{survie(i-1)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&\qquad \times \hat{Prob}[\textit{survie(i-1)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&= \hat{Prob}[\textit{survie(i)} \mid \textit{survie(i-1)}] \\
&\qquad \times \hat{Prob}[\textit{survie(i-1)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&= (1 - \hat{q}_i^*) \times \hat{Prob}[\textit{survie(i-1)} \mid \textit{survie(i-2)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&\qquad \times \hat{Prob}[\textit{survie(i-2)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&= (1 - \hat{q}_i)^* \times \hat{Prob}[\textit{survie(i-1)} \mid \textit{survie(i-2)}] \\
&\qquad \times \hat{Prob}[\textit{survie(i-2)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&= (1 - \hat{q}_i^*)(1 - \hat{q}_{i-1}^*) \\
&\qquad \times \hat{Prob}[\textit{survie(i-2)} \mid \textit{survie(i-3)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&\qquad \times \hat{Prob}[\textit{survie(i-3)} \text{ et } \ldots \text{ et } \textit{survie(1)}] \\
&= \cdots \\
&= (1 - \hat{q}_i^*)(1 - \hat{q}_{i-1}^*) \cdots (1 - \hat{q}_2^*)(1 - \hat{q}_1^*)
\end{aligned}$$

Soit donc :

$$\begin{aligned}
Prob[\text{survie au cours du } i^{\text{ième}} \text{ intervalle}] &= \hat{S}_i = \prod_{j=1}^{i} \hat{\pi}_j^* \\
&= \hat{\pi}_i^* \hat{S}_{i-1}
\end{aligned}$$

avec naturellement $\hat{S}_0 = 1$.

Pour trouver l'écart-type de cet estimateur on reprend la démarche utilisée pour dériver la formule de Greenwood donnant la variance de l'estimateur de Kaplan-Meier [28]. Ici, le point de départ est de noter que comme $\hat{q}_i^*$ est l'estimateur d'une proportion, son écart-type est donné par :

$$s(\hat{q}_i^*) = \sqrt{\frac{\hat{q}_i^*(1 - \hat{q}_i)^*}{n_i^*}} = \sqrt{\frac{\hat{q}_i^* \hat{\pi}_i^*}{n_i^*}},$$

et celui de la survie estimée sur le i$^{\text{ième}}$ intervalle est :

$$s(\hat{S}_i) = \hat{S}_i \sqrt{\sum_{j=1}^{i} \frac{\hat{q}_j^*}{n_j^* \hat{\pi}_j^*}}$$

28. Cf. l'équation (2.3).

Intervalle	n_i (a)	d_i (b)	c_i (c)	Effectif risqué moyen (d)=(a)-(c)/2	$1 - p_i$ =1-(b)/(d)	$\hat{S}$
[0,3[	20	2	0	20	0.90	$\hat{S}(0) = 1.00$
[3,6[	18	5	0	18	0.72	$\hat{S}(3) = 0.65$
[6,9[	13	0	0	13	1	$\hat{S}(6) = 0.65$
[9,12[	13	3	0	13	0.77	$\hat{S}(9) = 0.50$
[12,15[	10	2	2	9	0.78	$\hat{S}(12) = 0.39$
[15,18[	6	0	2	5	1	$\hat{S}(15) = 0.39$
[18,21[	4	2	0	4	0.50	$\hat{S}(18) = 0.19$
[21,∞[	2	0	2	1	1	$\hat{S}(21) = 0.19$

TABLE 2.15 – Exemple de calculs pour une table de survie, méthode actuarielle

Un exemple

Cet exemple utilise des données regroupées : on ne connaît pas pour chaque individu la date exacte de l'évènement ou de la censure mais seulement son appartenance à un intervalle de temps correspondant ici à un découpage en trimestres d'informations mensuelles. Les trois premières colonnes du tableau 2.15 correspondent aux informations de départ. Ainsi, pour la période allant du $12^{\text{ième}}$ mois inclus au $15^{\text{ième}}$ mois exclu, 10 individus sont risqués au début de l'intervalle, 2 vont connaître l'évènement étudié et 2 sont censurés.
Notez bien que les informations afférentes à un intervalle de temps donné sont utilisées pour construire l'estimation de la survie pour le début de l'intervalle suivant. Si on fait l'hypothèse d'une distribution uniforme des survenues d'évènements et des censures au cours de chaque intervalle de temps, alors la représentation graphique ne sera plus une fonction en escalier mais doit simplement relier entre eux les diverses survies estimées comme le montre le graphe 2.6.

Remarques :

— Lorsque d_i est nul alors la probabilité conditionnelle estimée sur le $i^{\text{ième}}$ intervalle est nulle. Ceci est naturellement techniquement exact mais peut être en pratique complètement irréaliste et montre que le choix des intervalles de temps a un impact sur les résultats de l'analyse.
— Si l'amplitude des intervalles tend vers zéro alors les estimations données par la méthode actuarielle tendent vers celles de l'estimateur de Kaplan-Meier. Pour cette raison ce dernier est aussi appelé *product-limit estimator*
— Une des raisons pour lesquelles les deux estimateurs diffèrent provient de la non similitude du traitement des données censurées.
— Par ailleurs l'estimateur KM donne une estimation de la survie pour tous les temps d'évènements observés et l'estimateur reste constant entre deux temps d'évènement observés, alors que la méthode actuarielle donne des estimations pour les durées correspondant aux bornes supérieures des intervalles (avec naturellement toujours $\hat{S}(0) = 1$).
— Avec la méthode actuarielle on peut encore estimer les fonctions de risque instantané $h(t)$ et de densité $f(t)$. Si on note $\bar{t}_{i-1,i}$ la durée correspondant au milieu du $i^{\text{ième}}$ intervalle, on utilise habituellement les expressions suivantes :

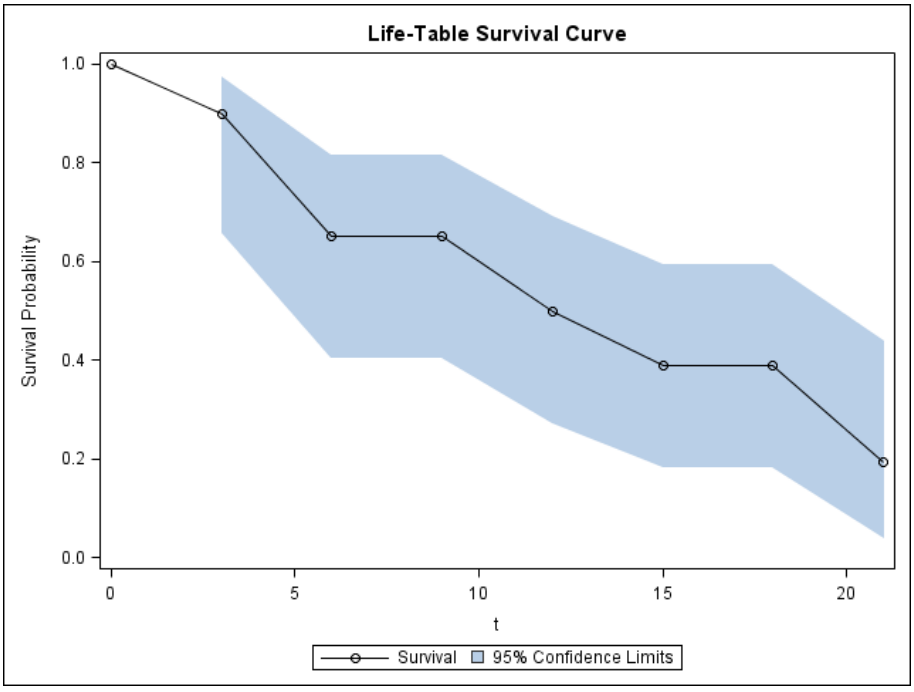


FIGURE 2.6 – Survie estimée, méthode actuarielle

$$\hat{f}(\bar{t}_{i-1,i}) = \frac{\hat{S}(t_i)\hat{q}_i^*}{t_i - t_{i-1}} = \frac{\hat{S}(t_i)\frac{d_i}{n_i^*}}{t_i - t_{i-1}}, \text{ et}$$

$$\hat{h}(\bar{t}_{i-1,i}) = \frac{\hat{q}_i^*}{(t_i - t_{i-1})(1 - \frac{\hat{q}_i^*}{2})}.$$

En pratique, sur petits échantillons ces estimateurs ne sont pas particulièrement bons.

2.11 PROC LIFETEST

On ne précise ici que les principales commandes et options ainsi que la syntaxe minimale. Pour plus de détails voyez l'aide de la proc.

PROC LIFETEST <options> ;

TIME variable <*censor(list)> ;

BY variables ;

FREQ variables ;

STRATA variable <(list)> <variable <(list)» . . . <variable <(list)» </options> ;

TEST variables ;

A l'exception de la commande TIME toutes les autres sont optionnelles. Cette commande requiert le nom d'une variable dont les valeurs sont les temps de survie ; censor est une variable indicatrice de la censure et list donne les valeurs de censor pour lesquelles il y a censure. Les observations à valeurs manquantes dans variable où censor ne sont pas prises en compte.

— **TIME** variable <*censor(list)> : variable contient les durées de survie à analyser. Elle peut être suivie par *censor(list), où censor est une indicatrice des survies censurées à droite. La liste (list) précise les valeurs pour lesquelles la censure est effective. Par exemple :
TIME duree ;
signifie qu'aucune valeur n'est censurée : tous les temps d'évènement sont observés.
TIME duree*cens(0) ;
les observations de duree pour lesquelles celles de cens valent 0 sont censurées
TIME duree*cens(1,5) ;
les observations de duree pour lesquelles celles de cens valent 1 ou 5 sont censurées.

— **BY** variables : constitue des sous-échantillons d'observations de temps de survie pour chacune des modalités prises par la (les) variable(s) spécifiées dans cette commande. Par exemple si sexe est une variable à deux modalités (1=homme, 2=femme), on réalisera une analyse de survie pour chacun des deux sexes avec la commande "**BY** sexe ;". A la différence de **STRATA**, aucun test d'homogénéité des distributions (du type log-rank ou Wilcoxon) n'est effectué. Lorsqu'on utilise cette commande **BY**, le fichier de données est supposé être trié préalablement selon les valeurs des variables en question au moyen de l'application de PROC SORT (il peut exister des exceptions à cette règle, voir la documentation SAS).

— **FREQ** variable : variable contient la fréquence de survenue de chacune des observations. Ainsi une observation donnée est considérée comme apparaissant n fois si n est la valeur de l'observation correspondante dans variable. Par exemple, supposez que l'on ait les observations suivantes :

```
data obs ;
input duree cens ;
cards ;
5 1
8 1
8 1
3 0
;
```

on pourrait faire :

```
data obs ;
input duree cens eff ;
cards ;
5 1 1
8 1 2
3 0 1
;
```

et utiliser l'instruction "`FREQ eff;`" dans l'appel de LIFEREG. Lorsque la variable contient des fréquences non entières, elles sont tronquées à l'entier inférieur. Une conséquence est que si la fréquence est inférieure à 1 alors l'observation concernée n'est pas utilisée.

— **STRATA** variable<(list)> : constitue des sous-échantillons d'observations de temps de survie pour chacune des modalités prises par la (les) variable(s) spécifiées dans cette commande. Cette commande implique l'estimation des probabilités de survie pour chacun des échantillons (à l'image de **BY** qui est plus efficace de ce point de vue) mais réalise également des tests d'homogénéité (log-rank et Wilcoxon par défaut) entre les différentes strates. Si (liste) n'est pas spécifiée alors chaque modalité qui n'est pas une valeur manquante de la variable définit une strate. Il est possible de créer une strate sur les valeurs manquantes en spécifiant l'option **MISSING**. La présence d'une liste créée des intervalles qui chacun vont définir une strate. Par exemple
STRATA age(16 20 30 40 50 65) ; ou **STRATA** age(16, 20, 30, 40, 50, 65) ;
produit les intervalles
$]-\infty, 16[, [16, 20[, [20, 30[, [30, 40[, [40, 50[, [50, 65[, [65, +\infty[$

La syntaxe d'indication de la liste est assez souple. Ainsi on peut obtenir les mêmes résultats que dans l'exemple ci-dessus avec :
STRATA age(16, 20 to 50 by 10, 65) ;
La présence de plusieurs variables dans **STRATA** génère des strates croisant les modalités spécifiées, ainsi :
STRATA age(16, 20 to 50 by 10, 65) sexe ;
va produire 14 strates avec estimation des probabilités de survie et tests d'homogénéité associés : les 7 strates correspondant au découpage en 7 intervalles de la variables age vont être créées pour les hommes d'une part et les femmes d'autre part. Le nombre de sous-échantillons considérés peut donc devenir rapidement élevé.
Parmi les options associées à cette commande, on trouve notamment :

- **GROUP=** qui permet de spécifier une variable dont les modalités définissent des strates sur lesquelles on veut réaliser un test d'homogénéité. Cependant les tests sont stratifiés en fonction des strates identifiées par la ou les variables indiquées au niveau de la commande **STRATA**.
- **ADJUST=** spécifie la méthode d'ajustement à employer en cas de réalisations de tests d'égalité des survies sur plusieurs couples de variables. On peut par exemple spécifier ADJUST=BONFERRONI (ou ADJUST=BON), ADJUST=SIDAK, etc. . ..
- **DIF=**. Avec DIFF=ALL tous les couples possibles sont comparés. Avec DIF=CONTROL('zzz'), on utilise la courbe de survie identifiée par 'zzz' comme référence pour les comparaisons.
- **TEST=** indique le ou les tests à mettre en oeuvre : TEST=LOGRANK demande la construction du test de Logrank. On obtient Wilcoxon avec TEST=WILCOXON). Parmi les autres tests, on peut citer PETO, TARONE, FLEMMING(p, q) ou p et q sont des valeurs positives ou nulles, etc. . .. Avec TEST=ALL on génère tous les tests actuellement disponibles (FLEMING(1,0) étant alors calculé par défaut).

— **TEST** variables : sert à tester l'influence de variables continues sur les probabilités de survie au moyen de tests de rang. En plus des tests de significativité individuelle, les résultats d'une procédure de sélection de type Forward sont également donnés.

Les principales options pouvant apparaître dans l'appel de la procédure **LIFETEST** sont les suivantes (voir la documentation de SAS pour l'ensemble des options disponibles) :

— **METHOD**= donne la méthode à utiliser pour l'estimation.
  METHOD = PL ou KM pour Kaplan-Meier
  METHOD = ACT ou LIFE ou LT pour la méthode actuarielle
  Par défaut METHOD= PL.
— **OUTSURV**=fichier ou **OUTS**=fichier : nom d'une table qui contiendra notamment les estimateurs des fonctions de survie (variable "SURVIVAL") et les bornes inférieures et supérieures des intervalles de confiance au seuil alpha (respectivement, variables "SDF_LCL" et "SDF_UCL), une indicatrice des observations censurées ("_CENSOR_, valant 1 si censure, 0 sinon.)".
— **ALPHA**= seuil de risque à utiliser pour construire les intervalles de confiance.
— **ALPHAQT**= idem ci-dessus mais pour les IC des quartiles.
— **INTERVALS**= bornes des intervalles à considérer pour construire les tables de survie. Par défaut SAS les sélectionne automatiquement. Vous pouvez cependant les imposer. Ainsi :
  INTERVALS= 4 8 12 16 ou **INTERVALS**= 4 to 16 by 4
  Construira la table de survie pour les intervalles
  $$[0, 4[, [4, 8[, [8, 12[, [12, 16[, [16, \infty[$$
— **WIDTH**= valeur numérique spécifiant la largeur éventuellement désirée des intervalles à considérer pour la construction des tables de survie.
— **NINTERVAL**= nombre d'intervalles à construire pour les tables de survie. Par défaut **NINTERVAL**= 10. Cette option est ignorée si **WIDTH**= est utilisée (cette dernière étant elle-même ignorée si **INTERVALS**= est précisé).
— **MAXTIME**= spécifie la durée maximale à considérer dans les représentations graphiques. Ceci n'affecte que les graphiques, pas les estimations.
— **MISSING** permet aux valeurs manquantes d'une variable numérique ou au blanc d'une variable alphanumérique de définir une strate lorsqu'on emploie la commande **STRATA**. Attention à ne pas la confondre avec l'option MISSING qui peut apparaître dans la commande STRATA créant une strate pour les observations manquantes de la variable spécifiée dans GROUP= .
— **NOTABLE** supprime l'affichage des estimateurs des fonctions de survie. Les graphiques et les tests éventuels d'homogénéité sont réalisés. Cette option est utile lorsque le nombre d'observations est important.
— **PLOTS**= (type) : demande l'affichage graphique de la fonction définie par le type, le graphe étant réalisé pour chaque strate si STRATA est actif.

CENSORED ou C	:	graphe des observations censurées
SURVIVAL ou S	:	graphe des fonctions de survie estimées $\hat{S}(t)$ versus t
LOGSURV ou LS	:	graphe de $-\log[\hat{S}(t)]$ versus t
LOGLOGS ou LLS	:	graphe de $\log\{-\log[\hat{S}(t)]\}$ versus $\log(t)$
HAZARD ou H	:	graphe de la fonction de risque estimée $\hat{h}(t)$ versus t
PDF ou P	:	graphe de la fonction de densité estimée $\hat{f}(t)$ versus t (seulement pour les tables

on pourra par exemple utiliser :

PLOT=(s) ou **PLOT**=(s,h)

Des options sont disponibles, on peut ainsi faire apparaître le nombre d'individus à risque pour certaines durées via `ATRISK(liste de durées)`, les intervalles de confiance avec CL (IC ponctuels) ou CB=ALL ou EP ou HW (Bande de confiance de Nair ou de Hall-Wellner), NOCENSOR supprime les indications des temps censurés (seulement avec KM), STRATA=INDIVIDUAL ou OVERLAY ou PANEL gère le graphe des survies si plusieurs strates sont étudiées, TEST fait apparaître dans le graphique la valeur du ou des tests d'homogénéité spécifiés dans la commande **STRATA**.

Nous verrons l'utilité éventuelle des graphes de $-\log[\hat{S}(t)]$ versus t et de $\log\{-\log[\hat{S}(t)]\}$ versus $\log(t)$ dans le chapitre suivant.

Chapitre 3

L'approche paramétrique

Si on suppose une distribution particulière des temps de survie alors il est possible d'introduire des variables explicatives dans la modélisation du risque. Par ailleurs si la distribution postulée est correcte alors les estimateurs obtenus sont plus efficients que les estimateurs non paramétriques. Comme, quelle que soit la distribution considérée, la fonction de survie est toujours non croissante, à réalisations dans $[0,1]$, il va être difficile de distinguer deux de ces fonctions associées à des distributions différentes, leur allure générale étant similaire. Pour cette raison on préfère travailler sur la fonction de risque mieux à même de représenter les a-priori que l'on peut avoir sur le phénomène étudié. Par exemple, une courbe en "U" peut être appropriée lorsqu'on suit des populations humaine depuis la naissance : après des taux de mortalité en bas-âge qui peuvent être relativement élevés, on observe leur décroissance par la suite avant la reprise d'une hausse aux âges avancés. En médecine une fonction de risque décroissante est souvent spécifiée lorsqu'on étudie les patients atteints de cancers : juste après le diagnostic le taux de mortalité est relativement élevé puis il décroît sous l'influence des traitements et des guérisons. En ingénierie, une courbe de risque constante est couramment admise pour modéliser la durée de vie des éléments électroniques.

Un des risques liés à l'emploi d'une méthode paramétrique consiste bien évidemment en un choix erroné de la distribution supposée et il importe donc de chercher à s'assurer de la pertinence du choix effectué. Une première indication, comme nous allons le voir, peut être tirée de l'évolution attendue a priori de la fonction de risque : celle-ci peut être incompatible avec telle ou telle distribution. Il existe également des tests permettant d'aider l'utilisateur à sélectionner une distribution plutôt qu'une autre. De même un certain nombre de graphiques peuvent apporter des informations utiles même s'il ne faut pas en exagérer la portée.

On retiendra en outre que la procédure d'estimation des modèles de survie paramétriques sous SAS autorise aisément la prise en compte de censure à droite, à gauche, et par intervalle. La première correspond au cas couramment traité dans le chapitre précédent : l'évènement se produira à une date inconnue à laquelle correspond un temps d'événement supérieur à la durée observable. La deuxième survient lorsque l'on sait que l'événement s'est produit avant une durée connue mais on ignore exactement quand. La dernière apparaît lorsque la date de réalisation de l'événement n'est pas connue avec précision : on sait seulement qu'elle appartient à certain un intervalle de temps qui lui est connu.

Le point commun de tous les modèles paramétriques pouvant être estimés par la procédure LIFEREG de SAS est de supposer une hypothèse dite de temps de vie accéléré (*Accelerated Failure Time* ou AFT) traduisant le fait que si $S_i()$ et $S_j()$ sont les temps de survie afférents à deux individus i et j, alors il existe une constante ϕ_{ij} telle que $S_i(t) = S_j(\phi_{ij}t)$ pour tout t. Allison (1998) donne comme exemple de cette configuration la relation souvent affirmée selon laquelle une année de la vie d'un chien équivaut à sept années de la vie d'un homme.
Ces modèles AFT se distinguent notamment des modèles dits à risque proportionnel (*Proportional Hazard* ou PH) qui sont caractéristiques des modélisations semi-paramétriques et notamment de la plus utilisée d'entre elles, le modèle de Cox qui fera l'objet du chapitre suivant. Les développements qui suivent précisent ces deux cadres d'analyse dont il importe de comprendre la signification.

3.1 Les modèles AFT et les modèles PH

3.1.1 Les Modèles à temps de vie accélérée

L'équation de base des modèles AFT

Dans ce type de modèle on explique le temps de survenu de l'événement d'intérêt. Un ensemble de k explicatives caractéristiques de chaque individu peut être mobilisé pour cette explication : $T_i = f(x_{i1}, x_{i2}, \ldots, x_{ik})$. Comme fonction de lien, on va retenir une forme usuelle : $f()$ est supposée linéaire. La seule précaution à prendre concerne le fait que le temps de survenue T_i est strictement positif. On intègre aisément cette condition via une transformation de l'expliquée, pour arriver à l'équation de base des modèles AFT :

$$\log(T_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + b\, u_i \tag{3.1}$$
$$= a_i + b\, u_i \tag{3.2}$$

Grâce à cette transformation, aucune contrainte particulière, ni de taille, ni de signe, n'a besoin d'être imposée sur les coefficients β. Notez également dans cette formulation usuelle des modèles AFT, que les caractéristiques des individus affectent uniquement le paramètre a_i, le paramètre b est supposé constant entre les individus [1].

Pour ce qui concerne le choix des distributions possibles pour u, les modèles à temps de vie accéléré retiennent des variables aléatoires obéissant au modèle dit *position-échelle* : Soit deux aléatoires réelles telles que $Y = a + bY_0$, $a \in \mathfrak{R}$, $b \in \mathfrak{R}$, $b > 0$ alors l'ensemble des lois de Y constitue un modèle position-échelle généré par Y_0, où a est le paramètre de position (location parameter) et b le paramètre d'échelle (scale parameter). Souvent Y_0 est l'aléatoire standardisée. Par exemple, si $Y_0 \sim \mathcal{N}(0,1)$ alors $\{Y \sim \mathcal{N}(\mu, \sigma^2)\}$ est engendré par Y_0 avec $Y = \mu + \sigma Y_0$ est . Si F_Y et F_{Y_0} sont les fonctions de répartition respectives de Y et Y_0 et y une réalisation de Y, on a :

$$F_Y(y) = F_{Y_0}(\frac{y-a}{b}) \tag{3.3}$$

ou bien encore, en termes de survie :

$$S_Y(y) = S_{Y_0}(\frac{y-a}{b}) \tag{3.4}$$

1. On peut imaginer d'introduire de l'hétérogénéité en faisant dépendre b de ces caractéristiques individuelles, $b = b_i = b(x_{i1}, \ldots, x_{ik})$. Dans SAS, cette possibilité n'est pas offerte en standard

Par ailleurs, si f_Y et f_{Y_0} sont leurs fonctions de densité, alors [2] :

$$f_Y(y) = \frac{1}{b} f_{Y_0}(\frac{y-a}{b}) \tag{3.5}$$

Vous aurez évidemment reconnu dans $Y_0 = \frac{Y-a}{b}$ la forme standardisée de Y pour laquelle le paramètre de position est égal à 0 et le paramètre d'échelle vaut 1. Un des intérêts pratiques des variables aléatoires de type position-échelle est que l'on peut, à condition de connaître les deux paramètres c et b faire des calculs qui impliquent F_Y ou f_Y avec uniquement des résultats afférents à la répartition et/ou la densité de la variable standardisée de la famille [3].

En quoi le temps est-il accéléré ?

Pour comprendre la dénomination de ces modèles, nous allons prendre un individu de référence pour lequel toutes les explicatives sauf la constante sont nulles. L'indice 0 servira à repérer cette référence. En reprenant l'équation de base des modèles AFT, on a évidemment :

$$Y_0 = \log(T_0) = \beta_0 + b\,u, \text{ soit encore,} \tag{3.6}$$

$$T_0 = \exp(\beta_0 + b\,u), \text{ et} \tag{3.7}$$

$$S_0(t) = Pr[T_0 > t] = Pr[u > \frac{\log(t) - \beta_0}{b}] \tag{3.8}$$

Pour un autre individu i, il vient :

$$Y = \log(T) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + b\,u, \tag{3.9}$$

$$T = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + b\,u) \tag{3.10}$$

$$\begin{aligned} S(t) &= Pr[T > t] = Pr[u > \frac{\log(t) - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}}{b}] \\ &= S_0\left(t \exp(-\beta_1 x_{i1} - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})\right) \\ &= S_0\left(\phi_i t\right) \end{aligned} \tag{3.11}$$

où $\phi_i = \exp(-\beta_1 x_{i1} - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})$, est un coefficient de proportionnalité positif et constant si les valeurs des explicatives sont invariantes dans le temps. Dans ce cas, la survie à une durée t pour un individu quelconque est égale à la survie observée à un autre durée donnée par $\phi_i t$ pour l'individu de référence. Pour reprendre l'exemple de l'homme et du chien, si le premier est en référence alors pour le chien, $\phi_i = 7$: la survie à 4 ans d'un chien serait égale à celle d'un homme de 28 ans. Tout se passe comme si le temps était accéléré pour le chien. En résumé pour un individu quelconque,

— Si $\phi_i > 1$, le temps s'accélère pour lui relativement à celui du référant,

— Si $\phi_i = 1$, le temps s'écoule à la même vitesse pour les deux individus,

2. L'obtention de (3.3) est immédiate : $F_Y(y) \equiv Prob[Y \leq y] = Prob[Y_0 \leq \frac{y-c}{b}] \equiv F_{Y_0}(\frac{y-c}{b})$. Celle de (3.5) est encore plus rapide si on se souvient que $f(y) = \delta F(y)/\delta y$. Il suffit donc de dériver (3.3) à gauche et à droite pour obtenir le résultat affiché.

3. Pour prendre un exemple connu, pensez à la recherche des valeurs critiques pour une gaussienne quelconque alors que les tables publiées ne concernent que la gaussienne standardisée.

— Si $\phi_i < 1$, le temps est décéléré pour l'individu i relativement à celui qui caractérise l'individu pris en référence.

Notez que cette dilatation du temps s'applique aux quantiles. Ainsi, si t_{M_0} est le temps médian pour la population de référence, i.e. $S_0(t_{M_0}) = 0.50$, alors pour l'individu i, $S(\phi_i\, t_{M_0}) = 0.50$. Pour reprendre l'exemple précédent, si l'âge médian au décès des hommes était de 70 ans, celui des chiens serait de 10 ans.
Compte tenu de l'équation de définition de ce paramètre de proportionnalité, il est évident que l'accélération où la décélération du temps dépend des caractéristiques des individus concernés via la valeur des explicatives $x_{i1}, x_{i2}, \ldots, x_{ik}$ et de leurs coefficients.

On va considérer un exemple simple pour illustrer les points précédents : Supposons que l'on soit en présence d'un modèle ne possédant, outre un terme constant, qu'une seule explicative codée 0 où 1 avec $S = 1$ si i est un homme et 0 sinon : les femmes sont donc ici utilisées comme référence. L'équation de ce modèle est alors :

$$Y = \log(T) = \beta_0 + \beta_1 S + b\, u \tag{3.12}$$

et la survie des femmes :

$$S_0(t) = Pr[u > \frac{\log(t) - \beta_0}{b}] \tag{3.13}$$

alors que celle des hommes devient :

$$S(t) = Pr[u > \frac{\log(t) - \beta_0 - \beta_1}{b}],$$

qui, si on l'exprime au moyen de la fonction de survie de femmes, s'écrit encore

$$S(t) = S_0\,(\exp(-\beta_1) \times t) \tag{3.14}$$

Avec ce codage binaire 0/1, il vient donc $\phi = e^{-\beta_1}$. En conséquence, si $\beta_1 > 0$ le temps décélère pour les hommes relativement aux femmes. Au contraire, si $\beta_1 < 0$ le temps d'événement s'accélère pour les hommes. Par exemple, pour $\beta_1 = 0.6931$, on a $\exp(\beta_1) = 2$: la survie des hommes à 20 ans est égale à celle des femmes à 10 ans, le temps des premiers a décéléré. Pour $\beta_1 = -0.6931$, on a $\exp(\beta_1) = 0.50$, et la survie des hommes à 20 ans serait égale à celle des femmes à 40 ans, le temps s'écoule deux fois plus rapidement pour les hommes.
Une autre formulation aidant à se rappeler de la relation existante entre le signe du coefficient d'une explicative et la vitesse d'écoulement du temps est la suivante : considérons une explicative affectée d'un coefficient positif. Si la valeur de cette variable augmente tourtes autres choses inchangées, alors d'après l'équation fondamentale des modèles AFT, le temps d'événement doit augmenter, ce qui signifie que la survie augmente, et équivaut à dire que le temps décélère. A l'inverse, avec un coefficient négatif, l'augmentation de la variable provoque une baisse des temps de survenue de l'événement, c'est à dire une diminution de la survie ce qui correspond à une accélération du temps.

L'interprétation des coefficients

Partant de l'équation de base des modèles AFT :

$$Y = \log(T) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + b\, u,$$

il est alors immédiat de calculer l'impact d'une variation dx_j de la valeur de la j^{ième} explicative toute autre variable inchangée :

$$dx_j \Rightarrow \frac{dT}{T} = \beta_j dx_j$$

ainsi, une variation d'un point de cette variable, $dx_j = 1$, provoque une variation relative de la durée de survenue de l'événement égale à β_j.

Remarques

1. la plupart des ajustements paramétriques de la survie ont comme expliquée le logarithme des temps de survie. Il est cependant possible dans la Proc LIFEREG d'expliquer T et non pas $\log T$ en utilisant l'option NOLOG. La transformation en logarithme est également désactivée si vous spécifiez pour les temps d'événement une distribution normale ou une distribution logistique. Rappelez-vous toutefois qu'il est possible alors que le modèle puisse générer des temps de survenus d'événement négatifs, ce qui est absurde. Il faut donc avoir de très bonnes raisons pour utiliser ces deux distributions.
2. Les distributions log-normale et de Weibull sont des exemples de distributions qui ne sont pas des modèles position-échelle. En revanche, prises en logarithme elles en font partie. On les retrouvera donc naturellement lorsqu'il s'agira de modéliser le logarithme des temps de survie.

Dans les modèles AFT, les explicatives affectent, via le paramètre de position, la durée de survenue d'un événement, et donc implicitement la fonction de survie. Une autre possibilité est de faire porter cette influence sur la fonction de risque. C'est ce que vont faire les modèles à risque proportionnels.

3.1.2 Les Modèles à risques proportionnels

Ici la spécification usuelle est :

$$h(t) = h_0(t)r(x) \tag{3.15}$$

où $h()$ est la fonction de risque. Elle est donc écrite comme le produit de deux fonctions, l'une $h_i0(t)$ étant dépendante du temps mais pas des caractéristiques individuelles, et l'autre, $r(x)$ ne dépendant pas du temps mais uniquement des caractéristiques des individus. A l'évidence, $h_0(t)$ est le risque d'un individu pour lequel $r(x) = 1$. Pour cette raison, $h_0()$ est également nommé risque de base.

L'explication du nom donné à ces modèles se comprend aisément si on considère le ratio de risque afférent à deux individus pour une durée t quelconque : on obtient une fonction indépendante du temps qui est de plus une constante si les valeurs des deux ensemble d'explicatives x_i et x_j sont elles-même invariantes :

$$\frac{h_i(t)}{h_j(t)} = \frac{r(x_i)}{r(x_j)} \tag{3.16}$$

Considérons ainsi la spécification suivante de $r(x_i)$, qui fait intervenir une exponentielle afin d'assurer la positivité de $h_i(t)$:

$$r(x) = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \tag{3.17}$$

il vient :

$$\frac{h_i(t)}{h_j(t)} = \exp\left[\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_k(x_{ik} - x_{jk})\right] \tag{3.18}$$

Bien que traitant de la fonction de risque, on peut remonter à la survie impliquée par un modèle PH. En effet, partant de la relation de base des modèles à risques proportionnels, $h(t) = h_0(t)r(x)$, en considérant l'égalité fondamentale

$$S(t|x) = exp\left[-H(t|x)\right] = exp\left[-\int_0^t h(v)dv\right],$$

et en utilisant la notation $S_0(t|x) = exp\left[-H_0(t|x)\right]$, on arrive immédiatement à :

$$S(t|x) = S_0(t|x)^{r(x)} \tag{3.19}$$

Remarques

1. Une caractéristique des modèles à risque proportionnels est que si $S_0(t)$ est dans une certaine famille de distribution paramétriques alors en général $S(t)$ n'est pas dans la même famille, ce qui est contraire à ce qu'on vérifie avec les modèles AFT. Ceci est une des raisons pour lesquelles les modèles paramétriques sont du type AFT plutôt que PH.
2. Un cas intéressant à considérer est la distribution de Weibull dont la densité s'écrit :

$$f_T(t|\alpha(x), \beta) = \frac{\beta}{\alpha(x)}\left(\frac{t}{\alpha(x)}\right)^{\beta-1} \exp\left[\left(\frac{t}{\alpha(x)}\right)^{\beta}\right], \tag{3.20}$$
$$\text{avec } t \geq 0, \beta > 0, \alpha(x) > 0.$$

   Si on considère $Y = log(T)$ alors on obtient une distribution de Gumbel pour l'aléatoire Y qui est de type position-échelle, avec :

$$f_Y(y|c(x), b) = \frac{1}{b}\exp\left[\frac{y - c(x)}{b} - \exp\left\{\frac{y - c(x)}{b}\right\}\right], \ -\infty < y < \infty, \tag{3.21}$$

   avec $c() = \log[\alpha(x)]$ et $b = \beta^{-1} > 0$.
   Par ailleurs, c'est également un modèle de type PH. En effet, la fonction de survie associée est :

$$S_T(t|x) = \exp\left[-\left(\frac{t}{\alpha(x)}\right)^{\beta}\right], \tag{3.22}$$

   et la fonction de risque :

$$h_T(t|x) = \frac{\beta}{\alpha(x)}\left[\frac{t}{\alpha(x)}\right]^{\beta-1} = \beta t^{\beta-1}\alpha(x)^{-\beta} \tag{3.23}$$

On note que cette dernière écriture est conforme à la spécification des modèles PH, puisque le risque total s'écrit bien comme le produit d'un terme ne dépendant que de t par un autre ne dépendant que des caractéristique x.

En d'autres termes, si on suppose que les temps de survie ont une distribution de Weibull, alors on est en présence d'un modèle qui est à la fois AFT et PH. La distribution de Weibull est d'ailleurs la seule à vérifier cette double appartenance.

3. Dans un modèle AFT l'équation estimée porte sur le logarithme des temps de survie. Dans un modèle PH elle porte sur la fonction de risque. De ce fait, les coefficients des mêmes explicatives estimés dans l'un ou l'autre cadre ne sont pas directement comparables. Supposons que l'on ait le même ensemble d'explicatives avec les coefficients β_{AFT} dans le premier et β_{PH} dans le second. Soit, en reprenant les notations précédentes, $\alpha(x) = \exp(\beta'_{AFT}x)$ d'une part, et $r(x) = \exp(\beta'_{PH}x)$ d'autre part. On pourrait s'attendre à ce qu'il soit de signe opposé : si une variable augmente le risque (son coefficient serait positif dans β_{PH}), alors elle devrait diminuer la survie (son coefficient serait négatif dans β_{AFT}. En fait la relation entre risque et survie est complexe [4] et il n'est pas rare, en pratique de ne pas observer cette inversion de signe. Le seul cas où le passage d'un ensemble de coefficient à l'autre est non ambigu est lorsque l'on travaille avec une distribution de Weibull [5]. Dans ce cas, l'expression de la fonction de risque (3.23) montre que $r(x) = \alpha(x)^{\frac{1}{\beta}}$ et donc $\beta'_{PH}x = -\frac{1}{\beta}\beta'_{AFT}x$, soit finalement : $\beta_{PH} = -\beta^{-1}\beta_{AFT}$. les coefficients sont proportionnels entre eux et de signe opposé.

3.2 Les principales modélisations AFT

Dans la classe des modèles AFT, les distributions exponentielle, Weibull, log-normale, log-logistique et gamma sont les plus couramment utilisées et implémentées dans la proc LIFEREG de SAS. Par la suite nous utiliserons pour décrire ces distributions les paramétrisations utilisées dans l'aide de la Proc LIFEREG.

3.2.1 La distribution exponentielle

Il s'agit du modèle le plus simple : on postule que le risque instantané est une constante :

$$h(t) = \alpha,\ t \geq 0,\ \alpha > 0. \tag{3.24}$$

Compte-tenu de la relation fondamentale entre la fonction de risque et la fonction de survie,

$$S(t) = \exp\left[-\int_0^t h(u)du\right],$$

il vient :

$$S(t) = \exp\left[-\int_0^t \alpha du\right] = e^{-\alpha t} \tag{3.25}$$

4. On rappelle que le passage du risque à la survie fait intervenir la densité.
5. Et ceci doit naturellement être relié à la remarque précédente.

La densité des temps de survie est alors donnée par :

$$f(t) = -\frac{dS}{dt} = \alpha e^{-\alpha t} \tag{3.26}$$

On reconnaît bien dans cette dernière expression la densité d'une exponentielle de paramètre α. Ainsi, une fonction de risque constante équivaut à une distribution exponentielle des durées d'événement. On vérifie immédiatement que la propriété AFT est vérifiée puisque pour deux individus i et j différents caractérisés par les paramètres constants $\alpha_i > 0$ et $\alpha_j > 0$, on a :

$$S_i(t) = e^{-\alpha_i t} = e^{-\phi_{ij}\alpha_j t} = S_j(\phi_{ij} t)$$

où ϕ_{ij} est la constante définie par $\phi_{ij} = \alpha_i/\alpha_j$.

Le risque ne variant pas avec le temps, il s'agit d'un processus sans mémoire. Pour cette raison, il est souvent utilisé en ingénierie pour modéliser notamment la durée de vie des composants électroniques. Afin d'illustrer cette propriété, on peut vérifier que pour deux durées t_1 et t_0 telles que $t_1 > t_0$, $Prob[T > t_1 | T > t_0] = Prob[T > t_1 - t_0]$. Ainsi, la probabilité qu'un composant fonctionne encore 3 ans, sachant qu'il a déjà fonctionné une année est simplement égale à la probabilité qu'il fonctionne deux années dès sa mise en fonction : il n'y a donc pas d'usure pendant la première année de fonctionnement [6].

3.2.2 La distribution de Weibull

La fonction de densité d'une variable aléatoire de Weibull s'écrit :

$$f(t) = \gamma \alpha t^{\gamma - 1} \exp(-\alpha t^{\gamma}) \text{ avec } \alpha > 0, \gamma > 0 \tag{3.27}$$

Les fonctions de survie et de risque associées sont respectivement :

$$S(t) = \exp(-\alpha t^{\gamma}) \tag{3.28}$$

$$h(t) = \gamma \alpha t^{\gamma - 1} \tag{3.29}$$

En conséquence, lorsque γ également appelé paramètre de forme, est supérieur à l'unité alors la fonction de risque est monotone croissante. Elle est monotone décroissante pour $\gamma < 1$ et constante si $\gamma = 1$. Cette flexibilité explique l'utilisation de cette distribution dès lors que l'on soupçonne une évolution monotone du risque avec la durée. On note également que la distribution exponentielle est un cas particulier de la Weibull obtenu avec $\gamma = 1$. Trois illustrations sont présentées dans le graphique 3.1

3.2.3 La distribution log-normale

La durée de vie T a une distribution log-normale si $Y = \log(T)$ est une distribution normale [7]. Si T est une log-normale et $Y = \log T$ est une normale d'espérance μ et de variance σ^2 alors sa

6. $Prob[T > t_1 | T > t_0] = \frac{Prob[T > t_1 \wedge T > t_0]}{Prob[T > t_0]} = \frac{Prob[T > t_1]}{Prob[T > t_0]} = \frac{S(t_1)}{S(t_0)} == \frac{e^{-\alpha t_1}}{e^{-\alpha t_0}} = S(t_1 - t_0)$

7. On voit immédiatement que $Y = \log T \sim N(\mu, \sigma^2) \Rightarrow T = \exp Y = e^{\mu + \sigma Y_0}$, où $Y_0 \sim N(0, 1)$. Le paramètre d'échelle de T est ainsi e^{μ}. Empiriquement, ce paramètre correspond à la moyenne géométrique des temps d'événements puisque $\hat{\mu} = 1/N \sum \log\{(t_i)\}$.

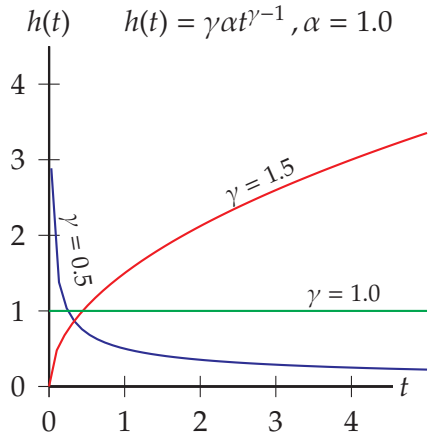


FIGURE 3.1 – Exemples de fonctions de risque avec la distribution de Weibull

densité est donnée par :

$$f_T(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{(\log\ t) - \mu}{\sigma}\right]^2\right),\ t > c, \sigma > 0. \tag{3.30}$$

Par ailleurs,

$$F_T(t) = \Phi\left(\frac{log(t) - \mu}{\sigma}\right), \tag{3.31}$$

$$S_T(t) = 1 - \Phi\left(\frac{log(t) - \mu}{\sigma}\right), \tag{3.32}$$

$$h_T(t) = \frac{\frac{1}{t\sigma}\phi\left[\frac{log(t)-\mu}{\sigma}\right]}{\Phi\left[-\frac{log(t)-\mu}{\sigma}\right]}, \tag{3.33}$$

$$H_T(t) = -\log\left(1 - \Phi\left(\frac{log(t) - \mu}{\sigma}\right)\right), t > 0, \sigma > 0. \tag{3.34}$$

où $\phi()$ et $\Phi()$ sont les fonctions de densité et de répartition répartition de la gaussienne standard. Par ailleurs,

$$E(T) = e^{\mu+\frac{1}{2}\sigma^2}, \text{et} \tag{3.35}$$

$$V(T) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} \tag{3.36}$$

L'intérêt de la log-normale est de pouvoir générer des fonctions de risque monotones mais aussi croissantes puis décroissantes en fonction de la valeur de σ, se démarquant ainsi de la Weibull. Elle a cependant un petit inconvénient : le calcul de $\Phi()$ implique le recours à des méthodes d'intégration numérique [8]. En conséquence, on lui préfère encore souvent la distribution log-

8. "petit" car compte-tenu de la puissance de calcul des processeurs actuels, les temps d'exécution ne sont plus très sensiblement dégradés par des appels à ces procédures d'intégration numériques.

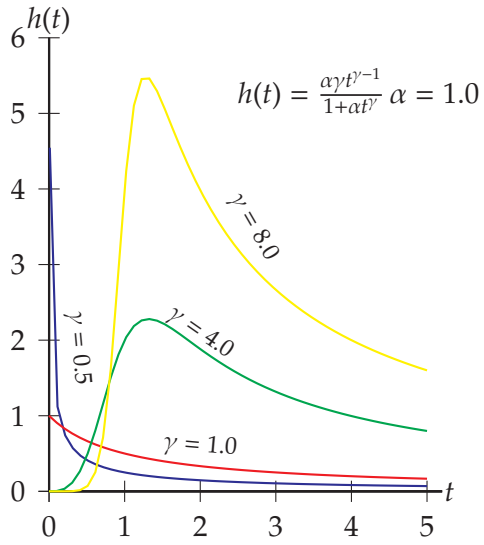


FIGURE 3.2 – Exemples de fonctions de risque sous distribution Log-logistique

logistique qui permet de générer des évolutions de risque similaires en ne faisant appel qu'à des fonctions élémentaires [9].

3.2.4 La distribution log-logistique

La distribution des temps d'événement est log-logistique si la densité de T est :

$$f(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})^2} \tag{3.37}$$

Les fonctions de survie et de risque associées sont respectivement :

$$S(t) = \frac{1}{1+\alpha t^{\gamma}} \tag{3.38}$$

$$h(t) = \frac{\alpha\gamma t^{\gamma-1}}{1+\alpha t^{\gamma}} \tag{3.39}$$

On rappelle que si T est une aléatoire log-logistique, alors $Y = \log T$ a une distribution logistique.

Des exemples de fonctions de survie obtenues pour diverses valeurs du paramètre de forme γ sont présentés dans le graphique 3.2. Rappelez-vous également que des évolutions similaires de la fonction de risque peuvent être obtenues avec la distribution log-normale.

3.2.5 La distribution Gamma généralisée

Cette distribution est intéressante car elle admet comme cas particulier les distributions Weibull et log-normale. Sa construction est cependant relativement compliquée et de plus la masse de

9. On retrouve ici quelque chose que vous connaissez : en économétrie des variables qualitatives, on sait qu'il est difficile de justifier l'emploi d'un Probit plutôt que d'un Logit : les deux ajustements conduisent généralement à des ajustements qualitativement identiques.

calculs nécessaires aux évaluations de sa densité est élevée ce qui peut affecter les temps de calcul. Plus grave, on peut rencontrer relativement souvent des difficultés de convergence dans le calcul des estimateurs lors de la maximisation de la log-vraisemblance.

La densité de cette gamma est donnée par :

$$f(t) = \frac{v}{t\sigma}\frac{|\delta|}{v\Gamma(\delta^{-2})}\left(\delta^{-2}v^{\delta}\right)^{\delta^{-2}}\exp\left(-v^{\delta}\delta^{-2}\right), v = \exp\left(\frac{\log t - \mu}{\sigma}\right) \tag{3.40}$$

$\Gamma()$ est la fonction gamma complète [10]. Le paramètre δ est un paramètre de forme [11]. L'intérêt de cette distribution dans les analyses de survie est double :

1. La Weibull et la log-normale sont des cas particuliers de la Gamma généralisée. La première est obtenue pour $\delta = 1$, la seconde lorsque $\delta = 0$. On retrouve la distribution exponentielle lorsque $delta = sigma = 1$. Ces propriétés peuvent être utiles pour fonder un test du rapport de vraisemblance afin d'aider au choix de la spécification de la distribution à retenir.
2. La distribution Gamma généralisée est donc capable de produire une fonction de risque semblable à celle des distributions énumérées au point précédent, mais de plus, elle peut aussi générer une fonction de risque en forme de U, décroissante puis croissante avec la durée, ce que ne peuvent faire les autres distributions vues jusqu'ici.

3.3 Estimation avec différents types de censure et tests sur les coefficients

La méthode d'estimation adaptée à l'estimation des modèles paramétriques est celle du maximum de vraisemblance. Celle-ci étant supposée connue, nous ne rappelons ici que des aspects très généraux.
Une fois le choix de la distribution effectué, l'écriture des diverses fonctions nécessaires aux calculs est connue. Pour des valeurs données des paramètres nous pouvons ainsi calculer la densité afférente à chaque individu pour lequel la durée reportée correspond à la réalisation de l'événement étudié. Soit $f_i(t_i)$ cette densité évaluée pour l'individu i qui a connu l'événement au temps t_i. L'indiciation de $f()$ par i rappelle que si des explicatives sont prises en compte alors les valeurs de ces explicatives devraient affecter l'évaluation de la densité : sur deux individus ayant le même temps d'événement mais des explicatives différentes on vérifiera généralement $f_i(t_i) \neq f_j(t_i)$.

Les observations correspondantes à une censure à droite sont celles pour lesquelles la durée d'événement est supérieure à la durée de censure. Leur vraisemblance est donc simplement la probabilité associée, soit, pour une durée t_i censurée à droite, $P[T_i > t_i] = S(t_i)$. Selon la même logique, une observation censurée à gauche, où l'on sait seulement que l'événement à eu lieu avant la durée t_i va intégrer la vraisemblance avec la probabilité correspondante, soit $P[T_i <= t_i] = F(t_i) = 1 - S(t_i)$. Enfin, si nous sommes en présence d'un individu censuré sur un intervalle, pour lequel on sait seulement que l'événement s'est réalisé entre t_{i_1} et t_{i_2}, avec naturellement $t_{i_1} < t_{i_2}$ la vraisemblance associée sera $P[t_{i_1} < T_i < t_{i_2}] = F[t_{i_2}] - F[t_{i_1}] = S[t_{i_1}] - S[t_{i_2}]$.

10. *Gamma*() est une généralisation de la fonction factorielle à des arguments réel et complexe. Pour les entiers, on a $\Gamma(n) = (n-1)!$
11. nommé "Shape" dans les sorties de Proc LIFEREG.

Ainsi, il est possible d'ajuster des échantillons présentant divers types de censures :

— Soit E l'ensemble des individus non censurés aux temps t_i,
— E_d celui des individus censurés à droite aux temps t_i,
— E_g celui des individus censurés à gauche aux temps t_i,
— E_i celui des individus censurés par intervalle aux temps t_{i_1} et t_{i_2},

la fonction de vraisemblance aura comme écriture générale :

$$L = \prod_{i \in E} f_i(t_i) \prod_{i \in E_d} S_i(t_i) \prod_{i \in E_g} F_i(t_i) \prod_{i \in E_i} S_i(t_{i_1}) - S_i(t_{i_2}) \tag{3.41}$$

On retrouve également tous les acquis de la théorie de l'estimation par le maximum de vraisemblance : s'il n'y a pas d'erreur de spécification, les estimateurs des paramètres sont asymptotiquement gaussiens, efficients. On sait estimer leur matrice de variance-covariance : la Proc LIFEREG utilise la méthode de Newton-Raphson qui évalue le hessien. A partir de là, il et possible de construire les tests usuels sur ces coefficients.

3.4 Choix d'une distribution et tests de spécification

On a compris, avec les développements précédents que l'une des premières difficultés rencontrées lors de l'ajustement d'un modèle paramétrique est celui du choix de la distribution sous laquelle vont se faire les estimations.Plusieurs méthodes peuvent être mises en oeuvre pour choisir et ou valider ce choix. Nous commencerons par la mise en oeuvre d'un test LRT de sélection d'une distribution, avant d'exposer les aides graphiques disponibles permettant d'orienter la décision vers une distribution appropriée. Si ces aides graphiques sont spécifiques aux méthodes de survie, en revanche, le test du rapport de vraisemblance vous est connu et donc son application doit se comprendre aisément. Dans tous les cas, rappelez-vous qu'une information a-priori sur l'évolution du risque en fonction de la durée constitue également un élément à ne pas négliger. Ainsi, si vous savez que le risque est constant, alors la distribution exponentielle mérite d'être envisagée.

3.4.1 Sélection au moyen du test de rapport de vraisemblance

Dès lors que le choix entre plusieurs distributions est possible, il est utile de pouvoir discriminer entre les diverses alternatives. Une solution, certes incomplète mais qui peut rendre service, est offerte par un test de type LRT. Pour mémoire, si l_{nc} et l_c sont les valeurs de la log-vraisemblance obtenues après estimation d'un modèle non contraint et d'un modèle contraint, alors la quantité $LRT = 2(l_{nc} - l_c)$ est distribuée selon un χ^2 à c degrés de liberté sous l'hypothèse nulle de validité des c contraintes imposées pour passer de l'un à l'autre. Ce test s'applique donc évidemment dans le cas où le modèle contraint est *emboîté* dans le non contraint [12].
Or, dans la section précédente, nous avons rencontré ce cas de figure à plusieurs reprises :

1. Le modèle exponentiel est un cas particulier du modèle Weibull. Il est obtenu lorsque le paramètre de forme γ est égal à l'unité. L'imposition de cette contrainte nous fait donc passer

12. On sait qu'un inconvénient de ce test est de rendre obligatoire l'estimation des deux modèles, contraint et non contraint, pour récupérer les valeurs des vraisemblances. Dans le cas présent, ceci n'est pas un obstacle majeur puisqu'il suffit de modifier un mot clef dans l'appel de la Proc LIFEREG pour changer de distribution.

d'une distribution de Weibull (modèle non contraint), à une distribution exponentielle (modèle contraint). Après ajustement de la même équation sous chacune de ces distributions, on récupère la vraisemblance estimée ($l_{nc} = \hat{l}_W$ et $l_c = \hat{l}_E$). On ne rejettera pas l'exponentielle en faveur de la Weibull si $LRT = 2(\hat{l}_W - \hat{l}_E) < \chi^2_\alpha(1)$, où $\chi^2_\alpha(1)$ est la valeur critique afférente au seuil de risque, α , choisi

2. Le modèle Weibull est lui-même un cas particulier de la Gamma généralisée. On passe de celui-ci à celui-là en contraignant le paramètre de forme, δ, de la Gamma à l'unité. En conséquence, On ne rejettera pas la Weibull en faveur de la Gamma si $LRT = 2(\hat{l}_G - \hat{l}_W) < \chi^2_\alpha(1)$.
3. Par ailleurs, lorsque $\delta = 0$ la Gamma généralisée dégénère en une log-normale. Ainsi on ne rejettera pas la log-normale en faveur de la Gamma si $LRT = 2(\hat{l}_G - \hat{l}_{LN}) < \chi^2_\alpha(1)$.
4. Enfin, si la Gamma peut dégénérer en Weibull, et que la Weibull peut elle-même dégénérer en exponentielle, on conçoit que l'exponentielle peut directement être dérivée de la Gamma. Il suffit simplement d'imposer deux contraintes pour réaliser le passage. Ainsi, on ne rejettera pas l'exponentielle en faveur de la Gamma si $LRT = 2(\hat{l}_G - \hat{l}_E) < \chi^2_\alpha(2)$.

Attention : le test LRT oppose deux modèles dont l'un est un cas particulier de l'autre. Il ne valide pour autant pas le modèle retenu. Par exemple, si on est amené à ne pas rejeter H0 en confrontant la distribution exponentielle et la Weibull, cela signifie simplement que le modèle le plus simple, ici l'exponentiel, est compte-tenu des données disponibles au moins aussi vraisemblable que le modèle Weibull. Si on rejette H0, alors l'exponentiel est moins vraisemblable que la Weibull. Mais, quelle que soit la décision, il est évidemment possible que les deux distributions soient en fait erronées, et que la vraie distribution soit toute autre.

3.4.2 Les aides graphiques

Celles-ci sont de deux ordres : aide au choix d'une distribution avant estimation d'une part, information sur le caractère approprié d'un ajustement paramétrique après estimation d'autre part.Il faut pourtant noter que malheureusement, en pratique, même si dans certains cas ils peuvent aider, le plus souvent les graphiques en question ne permettent pas de tirer une conclusion très assurée. Cette utilité limitée étant encore plus notable pour les aides graphiques avant estimation.

Aide au choix d'une distribution avant estimation

Elle concerne les modèles exponentiel, Weibull, log-normal et log-logistique. Pour les deux premiers les graphes sont aisément obtenus au moyen d'une option dans la Proc LIFETEST au moins pour deux des distributions vues précédemment..

— Pour le modèle exponentiel, la fonction de survie est donnée, en l'absence de variables explicatives, par $S(t) = e^{-\alpha t}$, soit encore $-\log S(t) = \alpha t$. Ainsi, si ce modèle est adapté, alors le graphe de $-\log S(t)$ en ordonnée sur t en abscisse doit être celui d'une droite passant par l'origine.
— Pour le modèle de Weibull, toujours en l'absence d'explicatives, l'équation de survie est $S(t) = \exp(-\alpha t^\gamma)$ d'où $\log[-\log S(t)] = \log\alpha + \gamma \log t$. Dans ces conditions, le graphe de

$\log[-\log S(t)]$ en ordonnée sur $\log t$ en abscisse doit donner une droite d'ordonnée à l'origine $\log \alpha$.

Comme nous l'avons vu au chapitre précédent, il est possible d'obtenir directement ces deux graphiques avec la Proc LIFETEST en spécifiant l'option PLOT=(ls) pour générer le premier, et PLOT=(lls) pour le second, où encore PLOT=(ls,lls) pour avoir les deux. Dans ces graphiques, l'estimateur de Kaplan-Meier $\hat{S}(t)$ de la survie se substitue naturellement à la valeur inconnue $S(t)$.

— Pour les distributions log-normale et log-logistique, l'obtention des graphes est un peu plus compliquée. Pour ces modèles, les fonctions de survie sont respectivement :

$$S_{LN}(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \tag{3.42}$$

$$S_{LL}(t) = \frac{1}{1 + \alpha t^{\gamma}}, \tag{3.43}$$

Soit :

$$\begin{aligned} \Phi^{-1}\left[1 - S_{LN}(t)\right] &= \frac{\log t - \mu}{\sigma}, \\ \frac{1 - S_{LL}(t)}{S_{LL}(t)} &= \alpha t^{\gamma}, \end{aligned}$$

et finalement :

$$\begin{aligned} \Phi^{-1}\left[1 - S_{LN}(t)\right] &= -\frac{\mu}{\sigma} + \frac{1}{\sigma}\log t, \\ \log\left[\frac{1 - S_{LL}(t)}{S_{LL}(t)}\right] &= \gamma \log \alpha + \gamma \log t. \end{aligned}$$

Ainsi, en appliquant à S(t) la transformation adapté à l'un ou l'autre modèle on devrait obtenir, avec cette transformation en ordonnée et $\log t$ en abscisse des graphes représentant des droites. La démarche à suivre est donc la suivante :

1. Un Appel de la Proc LIFETEST pour obtenir les estimateurs de Kaplan-Meier de $S(t)$ et sauvegarde de ceux-ci dans un fichier,
2. Une étape data pour crée les variables transformées [13],
3. Un appel à GPLOT pour obtenir les graphes.

Le squelette de cette construction serait donc :

```
Proc Lifetest data=...;
time duree*...;
outsurv=estim;
run;
Data estim;
```

13. On rappelle que *probit()* est sous SAS le nom de la fonction $\Phi^{-1}()$.

```
set estim;
lnormal=probit(1-survival);
logit=log(1-survival)/survival);
logt=log(duree);
run;
Proc gplot data=estim;
symbol value=none i=join;
plot lnormal*logt logit*logt;
run;
```

Aide au choix d'une distribution après estimation

Une limite forte de cette aide graphique avant estimation vient de ce que la démarche n'est plus pertinente dès lors que des explicatives sont présentes et affectent la survie : l'homogénéité que suppose ces graphiques n'est plus valide. La solution est de travailler avec les résidus du modèle estimé qui, si le modèle ajusté est satisfaisant doivent vérifier entre autres une hypothèse d'homogénéité [14].

La question est évidemment de définir les résidus pour un modèle de survie. Définissons les résidus comme les images d'une fonction ayant comme argument la variable expliquée, les explicatives et les paramètres du modèles. L'important est de définir la fonction de sorte à ce qu'elle renvoie un objet dont es propriétés sont connues si le modèle estimé est satisfaisant. Par exemple, ils devraient être en ce cas *i.i.d.* et de distribution connue. Soit $u_i = r(T_i, x_{i1}, x_{i2}, \ldots, , x_{ik}, \theta), i = 1 \ldots, n$. Si $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ, alors le résidu empirique est donné par $u_i = r(t_i, x_{i1}, x_{i2}, \ldots, , x_{ik}, \hat{\theta})$. Pour des valeurs suffisamment grande de n, ces résidus devraient se comporter comme des échantillons tirés dans la loi de u_i. Ce type de résidus est appelé *résidus généralisés* [15], et l'un des plus connu, particulièrement utile ici est le résidu de Cox-Snell définit par $u_i = H(T_i|x, \theta)$, soit encore, en fonction d'une relation fondamentale, comme $u_i = -\log S(T_i|x, \theta)$. On peut montrer que si la distribution choisie pour l'estimation du modèle paramétrique est satisfaisante, alors les résidus de Cox-Snell devraient être des aléatoires ayant une distribution exponentielle de paramètre égal à l'unité.

La preuve s'effectue en deux temps :

1. Si X est une aléatoire continue de fonction de répartition $F_X(x)$ alors l'aléatoire $U = F_X(x)$ possède une distribution uniforme sur $[0, 1]$ [16]. Évidemment, si $U = F_X(x)$ est une uniforme sur $[0, 1]$, alors c'est aussi le cas de $S_X(x) = 1 - F_X(x)$.

14. Pensez par exemple au modèle de régression linéaire $Y = X\beta + u$: l'espérance conditionnelle de Y égale à $X\beta$ est susceptible de se modifier pour chaque individu en fonction des valeurs des explicatives X. En revanche, dans le cadre des hypothèses usuelles tous les u_i sont centrés, homoscédastique et orthogonaux entre eux. On peut ainsi tester la validité d'une estimation OLS en travaillant sur les estimateurs $\hat{u}_i$ qui sont homogènes plutôt que sur $E[Y|X]$. Dans le cas présent, plutôt que de travailler sur $S(t)$ comme le fait l'aide graphique avant estimation, on préfère regarder si certaines propriétés sont valides sur l'équivalent des résidus $\hat{u}$. Naturellement ceci ne peut se faire qu'après estimation, lorsque l'on dispose des résidus empiriques en question.

15. "généralisé" pouvant être compris comme n'étant pas nécessairement une mesure d'un écart entre un y observé et un $\hat{y}$ calculé, mais comme un objet devant vérifier certaines propriétés si le modèle ajusté est satisfaisant.

16. $F_U(u) = P[U \leq u] = P[F_X(X) \leq u] = P[X \leq F_X^{-1}(u)] = F_X[F_X^{-1}(u)] = u$. Ce théorème est plus connu sous une formulation inversée : si U est une uniforme sur [0,1] alors $X = F_X^{-1}(U)$ a comme fonction de répartition $F_X()$. Il fonde la méthode dite de la transformation inverse : pour générer des pseudo-aléatoires ayant la distribution $F_X()$, il suffit de générer des réalisations u d'une uniforme sur [0,1] et de prendre $x = F_X^{-1}(u)$.

2. Si U est une uniforme sur $]0,1]$ alors $Y = -\log U$ est une exponentielle de paramètre égal à 1 [17].

Ainsi, en l'absence d'erreur de spécification les temps d'événement possèdent effectivement la distribution imposée pour l'estimation, alors $S(u)$ évaluée avec la distribution supposée est une uniforme et donc les résidus de Cox-Snell, définis par $-\log S(u)$ devraient être des exponentielles de paramètre $\alpha = 1$. En conséquence, toujours si le modèle sélectionné est adéquat, le comportement de $\hat{u}_i = -\log S(t_i|x_i, \hat{\theta})$ devrait donc approcher celui d'un échantillon d'exponentielles censurées. Nous savons alors que pour ce modèle, le graphe de $-\log(S(\hat{u}_i))$ en fonction de $\hat{u}_i$ doit être une droite passant par l'origine et de pente égale à 1.

Sous SAS la mise en oeuvre de cette aide graphique est simple :

1. Estimer le modèle sur les temps de survie avec la Proc LIFEREG en spécifiant une distribution parmi les 5 possibles (exponentielle, Weibull, log-normale, log-logistique, gamma généralisée) et récupérer dans un fichier les résidus de Cox-Snell en activant la commande `OUT=estim / CRESIDUAL=CS ou CRES=CS`, où `CS` est le nom de la variable qui contiendra ces résidus dans la table `estim`,
2. Appel de la Proc LIFETEST sur la table `estim` en spécifiant l'option `plot=(ls)` et la commande `Time CS*...`, la variable indicatrice de la censure étant la même que celle utilisée dans l'étape précédente lors de l'appel à la Proc LIFEREG.

un exemple

Afin d'illustrer la démarche précédente, nous allons utilser un fichier de données disponible sur le site de l'Institute for Digital Research and Education de UCLA [18] . L'objectif de ces données est d'étudier le temps de rechute d'anciens utilisateurs de drogue ayant subis deux types de traitement différents notamment selon leur durée (variable TREAT égale à 1 pour le programme long, égale à 0 pour le traitement court), et le site (variable SITE valant 0 pour l'un, et 1 pour l'autre) entre lesquels les patients ont été répartis aléatoirement. La variable AGE enregistre l'âge du patient lors de son entrée dans le programme de traitement. La variable NDRUTX donne le nombre de traitements auquel a été soumis le patient avant son incorporation dans l'un ou l'autre des deux traitements étudiés. Une variable, HERCO , précise le type de consommation dans les trois mois qui ont précédé son incorporation (HERCO=1 si consommation d'héroïne et de cocaïne, HERCO=2 si consommation d'héroïne ou de cocaïne, HERCO=3 si aucune des deux drogues n'a été prise pendant ces trois mois). Enfin, la variable TIME reporte le temps de retour à l'addiction et la variable CENSOR prend la valeur 1 si l'individu a effectivement rechuté après après le laps de temps donné par TIME, et vaut 0 si à cette durée il n'avait pas rechuté. On est donc dans cet exemple en présence d'une censure à droite.

Le modèle final retenu explique le temps de rechute par l'âge, le site, le nombre de traitements préalables, la durée du traitement, et incorpore un effet d'interaction entre l'âge et le site de

17. $F_Y(y) = P[Y \leq y] = P[-\log(U) \leq y] = P[U \geq e^{-y}] = 1 - F_U[e^{-y}] = 1 - e^{-y}$. En différenciant par rapport à y le premier et le dernier terme de cette suite d'égalités, il vient : $f_Y(y) = e^{-y}$, et on reconnaît là la densité d'une exponentielle de paramètre $\alpha = 1$.

18. Fichier "uis_small.sas7bdat", accessible à l'adresse suivante :
http ://www.ats.ucla.edu/stat/sas/seminars/sas_survival/uis_small.sas7bdat

Distribution	$\hat{l}_{max}$
Exponentielle	-983.49
Weibull	-980.74
Log-normale	-970.92
log-logistique	-961.76
Gamma	-967.50

TABLE 3.1 – maximum de la log-vraisemblance sous différentes ditributions

traitement. Sous Weibull, son estimation dans LIFEREG est donc commandée par le programme suivant :

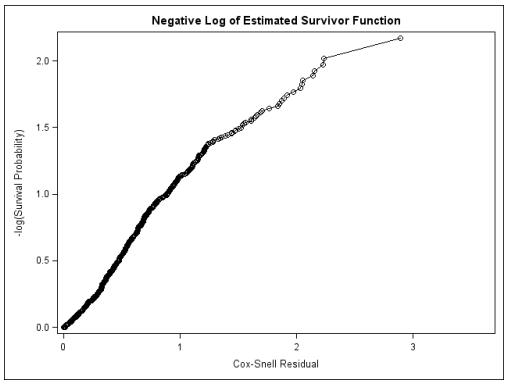
```
proc lifereg data=uis;
model time*censor(0) = age ndrugtx treat site age*site / D=WEIBULL;
output out=estim xbeta=index cres=cs;
run;
proc lifetest data=estim plots=(ls);
time cs*censor(0);
run;
```

A l'issue de la proc LIFETEST, on récupère le graphique (a) représenté dans la figure 3.3. En remplaçant `D=WEIBULL` dans le programme d'appel de LIFEREG par successivement `D=LOGNORMAL`, `D=LOGLOGISTIC` et `D=EXPONENTIAL`, la même proc LIFETEST va générer respectivement les graphiques (b), (c) et (d) du graphe 3.3. Finalement, il semble que les distributions log-normale et log-logistique conduisent à des représentations de droites plus satisfaisantes que l'exponentielle et la Weibull, et seraient donc préférées pour modéliser les temps étudiés dans cet exemple.

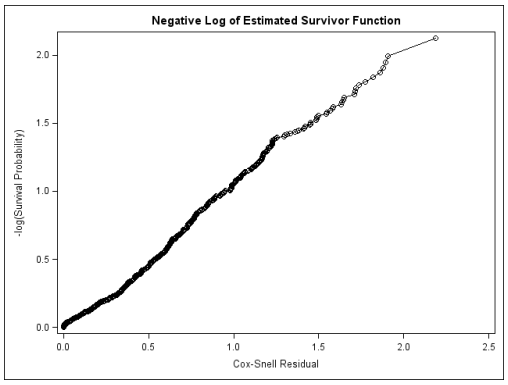
Nous pouvons également réaliser des tests LRT. Les estimations de l'équation précédente sous les cinq distributions disponibles donnent comme maxima pour la log-vraisemblance les valeurs présentées dans la table 3.1. Nous pouvons ainsi

— comparer exponentielle et Weibull :
$LRT = -2(-983.49 + 980.74) = 5.5 \overset{H0}{\sim} \chi^2(1)$
— comparer log-normale et gamma :
$LRT = -2(-970.92 + 967.50) = 6.84 \overset{H0}{\sim} \chi^2(1)$
— comparer exponentielle et gamma :
$LRT = -2(-983.49 + 967.50) = 31.98 \overset{H0}{\sim} \chi^2(2)$

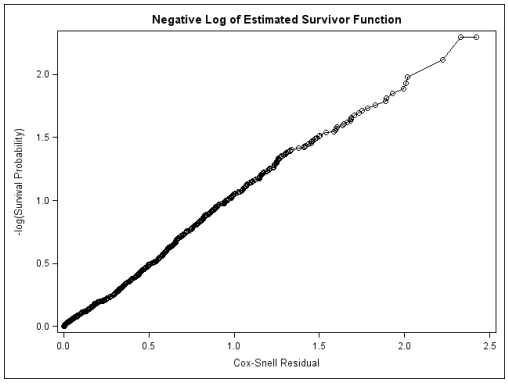
Avec des valeurs critiques à 5% de 3.84 et 5.99 respectivement pour les χ^2 à un et deux degrés de liberté, nous sommes amenés à rejeter l'exponentielle lorsqu'on l'oppose à la Weibull, à rejeter la log-normale et la Weibull si elles sont opposée à la Gamma. Au final, parmi toutes les distributions que nous pouvons opposer, ce test LRT serait donc favorable à une modélisation sous distribution gamma.



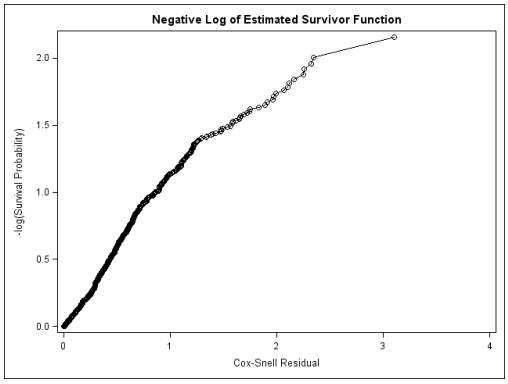
(a) Weibull



(b) Log-normale



(c) Log-logistique



(d) Exponentielle

FIGURE 3.3 – $-\log(survie)$ versus t sur résidus de Cox-Snell

3.5 estimation de fractiles sur les durées d'événement

Pour une ensemble d'explicatives, le p$^{\text{ème}}$ quantile y_p de $Y = \log T$ dans les modèles AFT est donné par :

$$y_p = X\beta + \sigma z_p \tag{3.44}$$

où z_p est le p$^{\text{ème}}$ quantile de la distribution standard de la famille considérée de densité $f_0(z)$.

L'estimateur de y_p est obtenu en remplaçant les paramètres inconnus par les valeurs des estimateurs du maximum de vraisemblance dans l'expression précédente. Sachant les caractéristiques d'un individu, on peut donc estimer par exemple, la durée t_{Med} telle qu'il ait 50% de chances de connaître l'événement entre 0 et t_{Med}. Par ailleurs, une estimation de l'écart-type d'un quantile estimé peut également être calculée au moyen de la méthode delta et la construction d'une forme quadratique faisant intervenir la matrice de variance-covariance des $\hat{\beta}$ selon la démarche habituelle.

3.6 Données censurées à gauche, à droite et par intervalle

La procédure LIFEREG autorise facilement la prise en compte des ces trois modes de censure. Nous avons vu précédemment comment, en leur présence, se construisait la fonction de vraisemblance . Le seul élément à préciser ici est celui de la syntaxe. Dans un premier point, on présente donc la structure attendue des données afin que LIFEREG identifie le cas de censure qui s'applique à chaque individu de l'échantillon. Dans un second temps, on profite des possibilités de cette procédure pour la mettre à contribution afin de réaliser l'estimation d'un modèle Tobit. Comme vous le savez, ce modèle est adapté à l'ajustement d'une régression sur une expliquée présentant une censure. Bien qu'il ne s'agisse pas d'un modèle de durée, cet exemple possède donc un intérêt propre. Par ailleurs, il illustre parfaitement une mise en application des règles de structuration des données afférentes à la Proc LIFEREG.

3.6.1 La structuration des données

Pour chacun des individu i non censurés, censurés à gauche ou censuré à droite, nous n'avons besoin que d'une seule durée t_i : soit l'événement s'est réalisé en t_i soit il s'est réalisé avant ou après t_i. Pour les individus censuré par intervalle, nous avons besoin de deux durées : les bornes basse et haute de l'intervalle, respectivement t_{i^-} et t_{i^+} signifiant que l'événement s'est produit entre ces deux durées.

Par défaut la censure est une censure à droite, ce qui explique que, s'il n'y a que ce type de censure, il suffit d'employer une syntaxe de la forme

```
model duree*cens(0)=...
```

LIFEREG comprend que les temps reportés dans la variable `durée` sont censurés à droite lorsque la variable `cens` prend la valeur 0. Dans le cas plus général, il faut une autre syntaxe. La solution adoptée dans LIFEREG est d'associer à chaque individu deux durées de vie L_i et U_i qui seront les i$^{\text{èmes}}$ observations des variables `L` et `U`. Leur interprétation est la suivante :

1. Si U_i est une valeur manquante, alors l'individu i est censuré à droite au temps L_i,
2. Si L_i est une valeur manquante, alors l'individu i est censuré à gauche au temps U_i,

individu	t_1	t_2
1	4.1	.
2	.	5.2
3	3.2	6.5
4	3.9	3.9
5	0.	4.3
6	5.6	1.7

TABLE 3.2 – Exemples de données associées à divers cas de censure

3. Si L_i et U_i ne sont pas des valeurs manquantes, et si $L_i < U_i$ alors l'individu i est censuré dans l'intervalle $[L_i, U_i]$,
4. Si $L_i = U_i$ et ne sont pas des valeurs manquantes, alors L_i (ou U_i) est un temps d'événement effectif : l'individu i n'est pas censuré.

   Par ailleurs deux autres règles s'appliquent :
5. Si $L_i = 0$, l'individu i n'est pas pris en compte dans l'estimation du modèle, cela parce qu'un temps d'événement doit être strictement positif,
6. Si $U_i < L_i$, l'individu n'est pas pris en compte dans les estimations en raison de l'incohérence des données qui lui sont associées : la borne basse de l'intervalle de censure est supérieure à la borne haute.

La syntaxe d'appel de la commande `model` devenant :

```
model (L U)=...
```

Soit par exemple les données du tableau 3.2. le premier individu a une temps d'événement censuré à gauche : celui-ci ne s'est pas réalisé, et s'il survient cela sera après plus de 4.1 unités de temps. Il s'agit donc d'une censure à droite. Le deuxième a une durée censurée à gauche : on sait seulement que pour lui l'événement s'est réalisé avant 5.2 unités de temps. Pour le troisième, l'événement s'est produit à une durée comprise entre 3.2 et 6.5 unités de temps mais on ne sait pas précisément quand. Pour le quatrième individu, l'événement s'est réalisé à une durée égale à 3.9. Les cinquième et sixième individus ne sont pas pris en compte dans les estimations du modèle. Le cinquième parce que la borne inférieure est nulle [19], le sixième car il présente une incohérence dans les bornes de l'intervalle si on spécifie que L contient les bornes inférieures et U les bornes supérieures de ces intervalles.

3.6.2 Estimation d'un modèle Tobit via LIFEREG

On peut utiliser cette capacité de la Prog LIFEREG à traiter les cas de censure précédents pour estimer un modèle de type Tobit [20]. Dans le cadre d'un modèle Tobit de type 1 avec double censure, on a une variable latente y^* et une variable observée y définies par :

19. Attention, dans le cas de ce cinquième individu, on pourrait l'interpréter comme une censure à gauche : l'événement s'est produit, on ne sait pas exactement quand entre le temps 0 et le temps 4.3. En toute logique cela n'est pas faux. Seulement LIFEREG ne retient que les temps ou les bornes d'intervalles de temps strictement positifs.
20. Depuis SAS 9, il est également possible d'employer la Proc QLIM

$$
\begin{aligned}
u &\sim N(0, \sigma^2) \text{ telle que } E[u|X] = 0, && (3.45)\\
y^* &= X\beta + u, && (3.46)\\
y_i &= \begin{cases} y_i^* \text{ si } a \leq y_i^* \leq b \\ a \text{ si } y_i^* \leq a \\ b \text{ si } b \leq y_i^* \end{cases} && (3.47)
\end{aligned}
$$

En d'autres termes, on observe la réalisation de y^* lorsqu'elle appartient à l'intervalle $]a, b[$, on observe seulement a (resp. b) si elle est inférieure (resp. supérieure) à a (resp. b). Dans ces conditions , y est une gaussienne censurée, $y \sim CN(X\beta, \sigma^2, a, b)$, et :

$$
\begin{aligned}
E[y_i^*|x_i] &= x_i\beta, \text{ où } x_i \text{ est la i}^{\text{ème}} \text{ ligne de } X && (3.48)\\
E[y_i|a < y_i < b, x_i] &= x_i\beta + \sigma \frac{\phi\left(\frac{a-x_i\beta}{\sigma}\right) - \phi\left(\frac{b-x_i\beta}{\sigma}\right)}{\Phi\left(\frac{b-x_i\beta}{\sigma}\right) - \Phi\left(\frac{a-x_i\beta}{\sigma}\right)} && (3.49)
\end{aligned}
$$

Le cas d'une censure à gauche, traité dans la recherche séminale de Tobin (1958) correspond à $y \sim CN(X\beta, \sigma^2, 0, +\infty)$: il s'agissait d'expliquer les acquisitions de biens durables dont les valeurs observées ne peuvent évidemment pas être négatives [21]. Pour changer, nous allons considérer le cas d'une censure à droite [22], correspondant à $y \sim CN(X\beta, \sigma^2, -\infty, b)$.

Pour cela nous allons reprendre un exemple donné sur le site de l'"Institute For Digital Research And Education" rattaché à UCLA [23]. On dispose pour 200 élèves des résultats d'un test d'aptitude aux études universitaires donnant un score compris, par construction entre 200 et 800 (variable `APT`). On veut expliquer ce score par deux autres mesures : un score de lecture (variable `READ`) et un score de math (variable `MATH`). Une dernière variable (variable `PROG`) ayant trois modalités indique le type de parcours suivi par chaque étudiant : `1=Academic, 2=General, 3=Vocational`. La censure provient du fait que tous les étudiants ayant répondus correctement lors du test à toutes les questions se voient attribués un score de 800, même si leur aptitude n'est pas égale. De même les étudiants ayant répondu incorrectement à toutes les questions du test reçoivent une score de 200, même si leur inaptitude n'est pas la même [24]. Sur la variable latente APT^*, le modèle ajusté est donc :

$$APT_i^* = \beta_0 + \beta_1 READ_i + \beta_2 MATH_i + \beta_3 Academic_i + \beta_4 General_i + u_i \quad (3.50)$$

où,

$$Academic_i = \begin{cases} 1 \text{ si PROG}_i = 1, \\ 0 \text{ sinon,} \end{cases}$$

21. Vous pouvez vous rafraîchir la mémoire à propos de ce modèle et de diverses variantes en consultant le polycopié de cours de Christophe Hurlin disponible à l'adresse suivante :
`http://www.univ-orleans.fr/deg/masters/ESA/CH/Qualitatif_Chapitre3.pdf`

22. Un exemple d'emploi de la LIFEREG avec censure à gauche est donné dans la documentation de cette procédure.

23. Pour plus de détails, voir `http://www.ats.ucla.edu/stat/sas/dae/tobit.htm`. L'exemple en question y est traité via la Proc QLIM citée auparavant.

24. Dans les données, la valeur minimale de `APT` est de 352 : aucun étudiant n'a reçu le score minimal de 200. En d'autres termes, si la censure à gauche est potentiellement présente, elle n'est pas effective sur les données utilisées, ce qui explique que nous limitions l'exercice à la seule censure à droite, 17 étudiants ayant obtenus un score de 800.

et,

$$General_i = \begin{cases} 1 \text{ si PROG}_i = 2, \\ 0 \text{ sinon,} \end{cases}$$

La variable observée, APT, est donc définie comme :

$$APT_i = \begin{cases} APT_i^* \text{ si } APT_i^* < 800, \\ 800 \text{ sinon} \end{cases} \tag{3.51}$$

Il suffit donc de caler les modalités de structuration des données conformément aux indications précédentes pour réaliser une estimation Tobit sur une censure à droite des paramètres de l'équation précédente. En nous conformant aux notations de la section précédente, il faut créer une configuration vérifiant $L = U = APT$ lorsque $APT < 800$, et $L = 800, U = .$ lorsque $APT = 800$. C'est ce que fait le programme suivant dans l'étape data, le rôle de L étant attribué à APT, celui de U à la variable *upper* [25] :

```
data tobit;
input id read math prog apt;
if apt = 800
then upper=.;
else upper=apt;
cards;
1 34 40 3 352
2 39 33 3 449
⋮
198 47 51 2 616
199 52 50 2 558
200 68 75 2 800
;
run;
```

Il suffit maintenant d'exécuter la Proc LIFEREG en spécifiant la distribution voulue, ici la gaussienne conformément à la pratique courante employée pour l'estimation des modèles Tobit. Notez qu'il serait également possible d'estimer le modèle sous la distribution logistique qui peut se révéler avantageuse dans certains cas de figure.

Important : rappelez-vous également que dans la Proc LIFEREG, en spécifiant l'une ou l'autre des distributions normale ou logistique, nous activons implicitement l'option NOLOG : l'expliquée n'est pas transformée.

Dans cet exemple, outre l'estimation des paramètres $\beta_1, \beta_2, \beta_3, \sigma^2$, nous allons également calculer les estimateurs des espérances des scores tronqués et non tronqués selon :

$$E[APT^*|x_i] = x_i\beta, \tag{3.52}$$

$$E[APT|APT < 800, x_i] = x_i\beta - \sigma\lambda\left(\frac{800 - x_i\beta}{\sigma}\right), \tag{3.53}$$

25. On ne reproduit ci-après que quelques lignes du fichier de données. Vous pouvez y accéder via l'adresse suivante :`http://www.ats.ucla.edu/stat/data/tobit.csv`

Analysis of Maximum Likelihood Parameter Estimates						
Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	163.42	30.41	28.88	<.0001
read		1	2.70	0.62	19.01	<.0001
math		1	5.91	0.71	69.43	<.0001
prog	1	1	46.14	13.72	11.30	<.0008
prog	2	1	33.43	12.96	6.66	0.0099
prog	3	0	0.0000	.	.	.
Scale		1	65.68	3.48		

TABLE 3.3 – Estimation d'un Probit avec censure à droite via LIFEREG

Dans la seconde équation, qui se déduit immédiatement de (3.49) lorsque $a = -\infty$, le terme $\lambda()$ est le ratio de Mill [26].
Pour cette raison, nous créons en sortie deux fichiers, le premier nommé `outest` ne contient qu'une observation, celle de s^2, l'estimateur de $sigma^2$, le second, `out`, contient les valeurs calculées de l'index $X\hat{\beta}$ pour chacun des 200 individus. Soit donc :

```
proc lifereg data=tobit outest=outest(keep=_scale_);
class prog;
model (apt upper) = read math prog / d=normal;
output out=out xbeta=index;
run;
```

Les résultats de l'estimation sont présentés dans la table 3.3
Enfin, les estimations des espérances sont produites par l'exécution du programme suivant :

```
data predict;
drop lambda _scale_ _prob_;
set out;
if _n_ = 1 then set outest;
lambda = pdf('NORMAL', (800-index)/_scale_)/ cdf('NORMAL', (800-index)/_scale_);
Predict = index-_scale_*lambda;
label index ='MOYENNE DE LA VARIABLE NON CENSUREE'
Predict = 'MOYENNE DE LA VARIABLE CENSUREE';
run;
proc print data=predict label;
format index Predict 7.2;
run;
```

Dans la table 3.4, nous présentons les résultats des calculs pour une dizaine d'élèves de l'échantillon.

26. $\lambda(z) \equiv \frac{\phi(z)}{\Phi(z)}$

Obs	read	math	prog	apt	upper	moyenne de APT non censurée	moyenne de APT censurée
191	47	43	2	567	567	577.98	577.89
192	65	63	2	800	.	744.83	721.80
193	44	48	2	666	666	599.46	599.21
194	63	69	2	800	.	774.92	737.37
195	57	60	1	727	727	718.22	704.71
196	44	49	2	539	539	605.37	605.05
197	50	50	2	594	594	627.47	626.64
198	47	51	2	616	616	625.29	624.53
199	52	50	2	558	558	632.87	631.83
200	68	75	2	800	.	823.90	755.39

TABLE 3.4 – Estimations des scores censurés et non censurés

3.7 PROC LIFEREG

Comme pour la Proc LIFETEST, nous ne présenterons ici que les principales commandes et options de LIFEREG avec leur syntaxe minimale. Pour plus de détails sur l'ensemble des possibilités, consultez l'aide de la Proc distribuée par SAS.

PROC LIFEREG
MODEL
BY
CLASS
OUTPUT <OUT=...>
WEIGHT

— **PROC LIFETEST** <options> ;
Appel de la procédure. Les principales options disponibles sont les suivantes :
— DATA=SAS-data-set
donne le nom du fichier de données sur lequel vont se faire les estimations.
— OUTEST= SAS-data-set
indique le nom du fichier qui contiendra certains résultats de l'estimation du modèle ajusté, notamment les estimateurs des coefficients, la valeur de la logvraisemblance, et, si l'option COVOUT est spécifiée, la matrice de variance-covariance estimée des coefficients.
— COVOUT
réclame l'écriture de la matrice de variance-covariance des estimateurs dans le fichier précisé par OUTEST.
— ORDER=DATA | FORMATTED | FREQ | INTERNAL
précise l'ordre de classement des modalités des variables utilisées dans l'instruction CLASS. Par défaut, c'est la spécification `ORDER=FORMATTED` qui est utilisée. Elle applique un ordre lexicographique sur sur les étiquettes de format si elles existent, sinon sur

les observations des variables alphanumériques et la relation "<" sur les valeurs des variables numériques. Avec ORDER=DATA, les modalités des variables sont rangées selon leur ordre d'apparition dans le fichier des données. Avec `ORDER=FREQ`, ces modalités sont rangées selon leurs effectifs, par ordre décroissant.
Par exemple, soit la variable genre binaire 0/1 sur laquelle on a appliqué le format 1 →"Femmes",0 →"Hommes". Avec l'emploi de `ORDER=FORMATTED`, la première catégorie de genre sera les femmes, la seconde les hommes. Si on ne met pas de format, alors la première catégorie de genre regroupera les hommes et la seconde les femmes.

— **CLASS** variables ;
Cette instruction s'applique s'applique à des variables dont les modalités définissent des catégories d'individus. Elle va créer des indicatrices de ces catégories qui pourront être intégrées à la liste des explicatives de l'équation ajustée définie par la commande `Model`. Si cette instruction est utilisée, elle doit apparaître avant la commande MODEL de façon à ce que les indicatrices soient créées avant d'être utilisées comme explicatives. La catégorie mise en base est toujours la dernière catégorie définie par l'option `ORDER`.

— **MODEL**
La commande MODEL est requise : elle précise notamment le nom de l'expliquée, les explicatives, la distribution utilisée. Elle peut prendre trois formes dont deux seulement ont été vues ici :
— MODEL expliquée*<censor(list)>=variables explicatives</options> ;
Sous cette forme, s'il y a censure, il s'agit d'une censure à droite. Dans ce cas, la variable censor donne l'information nécessaire : toutes ses observations égales aux valeurs précisées dans list signale que pour l'individu concerné, la variable expliquée est censurée.
— MODEL(lower,upper)=explicatives</options> ;
Avec cette syntaxe on autorise une censure à droite, à gauche ou par intervalle. L'interprétation s'effectue selon les règles vues précédemment et relatives aux valeurs des variables lower et upper.

Des effets croisés peuvent être aisément introduits, la liste des explicatives pouvant prendre la forme suivante : x_1 x_2 x_1*x_2, x_1 et/ou x_2 pouvant éventuellement être des variables de classification référencées dans la commande CLASS.

Parmi les options disponibles, on trouve :
— DISTRIBUTION=type, ou DIST=type, ou D=type,
avec comme choix possible de type de distribution pour les modèles AFT :
— exponentielle (EXPONENTIAL),
— Weibull (WEIBULL), la distribution utilisée par défaut,
— log-normale (LLNORMAL),
— log-logistique (LLOGISTIC), et
— gamma généralisée à trois paramètres (GAMMA).
Par défaut, la spécification de l'une de ces distributions ajuste le logarithme des temps d'événement. Il est possible d'empêcher cette transformation avec l'option NOLOG.
Deux autres distributions peuvent être spécifiées :
— normale (NORMAL),

— logistique (LOGISTIC),
qui ajustent les temps d'événement non transformés. Ainsi, les spécifications LLOGISTIC ou LLNORMAL avec l'option NOLOG sont équivalentes respectivement aux options NORMAL ou LOGISTIC.
— ALPHA=valeur. Précise le seuil de risque de première espèce qui doit être utilisée pour construire les intervalles de confiance sur les paramètres et la fonctions de survie estimée. Par défaut, ALPHA=5%.
— CORRB, réclame l'affichage de la matrice de corrélation des coefficients estimés.
— INITIAL=liste de valeurs. Permet d'imposer des valeurs initiales pour les paramètres à estimer, constante exclue, dans l'algorithme de maximisation de la log-vraisemblance. Cette option peut se révéler utile lorsque l'on rencontre des difficultés de convergence.
— INTERCEPT=valeur. Option qui permet d'initialiser le terme constant de la régression.
— NOINT. Supprime la constante du modèle ajusté.

— **OUTPUT**<OUT=SAS-data-set><mot clef=nom>...<mot clef=nom> ;
Cette commande réclame la création d'une table contenant toutes les variables inclues dans le fichier spécifié en entrée de la procédure plus un certain nombre de statistiques crées après l'estimation du modèle et dont la sélection s'opère par des mots-clés. Parmi ceux-ci :
— CDF=
nom d'une variable qui contiendra l'estimation de la fonction de répartition pour les temps d'événement effectivement observés,
— CONTROL=
nom d'une variable du fichier d'entrée qui, selon sa valeur, autorise ou non le calcul des quantiles estimés sur les individus concernés. Ces estimateurs n'apparaîtront que sur les individus pour lesquels elle vaut 1. En son absence, les quantiles seront estimés sur toutes les observations.
— CRESIDUAL= ou CRES=
nom d'une variable qui contiendra les valeurs des résidus de Cox-Snell, $-\log \hat{u}_i$.
— SRESIDUAL=
nom d'une variable contenant les résidus standardisés $u_i = \frac{y_i - x_i\hat{\beta}}{s}$.
— PREDICTED= ou P=
variable utilisée pour stocker les estimateurs des quantiles estimés sur les temps d'événement. Les individus pour lesquels l'expliquée est valeur manquante ne sont pas pris en compte dans l'étape d'estimation des paramètres du modèle. En revanche, on obtiendra les quantiles estimés pour ces mêmes individus, à condition naturellement que les explicatives soient renseignées.
— QUANTILES=liste ou QUANTILE=liste ou Q=liste
où liste précise les quantiles à estimer, les valeurs étant évidemment comprises entre 0 et 1 (exclus). Par exemple, Q=.25 .50 .75. Par défaut, Q=0.50.
— STD_ERR= ou STD=
nom d'une variable contenant les écarts-types des quantiles estimés.
— XBETA= nom d'une variable servant à stocker les estimations de l'index $X\hat{\beta}$

— **BY** variables ;
réclame l'ajustement du même modèle sur des sous-échantillons repérés par les modalités

des variables spécifiées. Rappelez-vous qu'en règle générale l'utilisation de cette commande doit être précédée par un PROC SORT, de sorte que le fichier d'entrée dans la LIFEREG soit déjà trié selon les variables en question.

— **WEIGHT variable ;**
précise le nom de la variable contenant les poids à donner aux individus dans la fonction de vraisemblance. Ces poids ne sont pas nécessairement des entiers, et les individus ayant un poids non positif ou valeur manquante ne sont pas utilisés pour l'estimation du modèle.

Chapitre 4

L'approche semi-paramétrique

Ce chapitre sera entièrement consacré à la présentation du modèle de Cox qui est l'approche la plus populaire dans l'analyse des modèles de survie. Comme nous allons le voir, ce modèle ne requiert pas la formulation d'une hypothèse de distribution des temps de survie, hypothèse qui est au coeur de l'approche paramétrique vue dans le chapitre précédent. Cependant l'estimation des paramètres du modèle et tout particulièrement des coefficients des variables explicatives passe par la maximisation d'une fonction de vraisemblance dite partielle.

Le coeur du modèle est la description du risque total comme le produit de deux éléments, le premier est un risque dit "de base", identique pour tous les individus, et qui ne dépend que du temps, alors que le second est fonction des caractéristiques des individus et plus généralement des explicatives retenues. L'estimation des coefficients de ces dernières, s'effectue en maximisant la vraisemblance obtenue en ne considérant qu'une partie de la vraisemblance totale, d'où le qualificatif de vraisemblance partielle. Les estimateurs obtenus seront naturellement moins efficients que ceux découlant de la maximisation de la vraisemblance complète mais cette perte d'efficience est toutefois contrebalancée par l'énorme avantage de ne pas avoir à spécifier de distribution particulière sur les temps de survie, ce qui doit accroître leur robustesse. Une fois obtenus les estimateurs des coefficients, il est possible de construire un estimateur non paramétrique du risque de base.

La conjugaison d'une estimation non paramétrique avec une technique de maximisation de vraisemblance explique que ce modèle soit qualifié de semi-paramétrique. Cette solution, proposée par Cox, est l'approche la plus couramment employée pour l'ajustement des modèles de survie. Pour donner une idée de sa popularité, Ryan & Woodall [1] dans un travail de 2005 proposent une liste des articles les plus cités dans le domaine de la statistique. Le premier [2] serait celui de Kaplan, E. L. & Meier, P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, pp. 457-481, et le second [3] serait celui de Cox, D. R. (1972) Regression models and life tables, *Journal of the Royal Statistical Society*, Series B, 34, pp. 187-220.

Le modèle de Cox est un modèle à risques proportionnels (PH) tel que nous l'avons défini dans le second chapitre. La fonction modélisée est la fonction de risque $h()$ alors que les modèles AFT

1. Ryan, Thomas P. & Woodall (2005), The Most-Cited Statistical Papers, Journal of Applied Statistics, Vol. 32, No. 5, 461-474, July 2005
2. Avec 25 869 citations
3. Avec 18 193 citations

que l'on estime avec la proc LIFEREG modélisent la fonction de survie $S()$. L'estimation du modèle de Cox sous SAS s'effectue via la proc PHREG.

4.1 Le modèle de Cox et son estimation

4.1.1 La fonction de vraisemblance partielle

Étant un modèle à risques proportionnels, il obéit à la spécification

$$h(t) = h_0(t)r(\mathbf{x}) \tag{4.1}$$

où $h_0(t) > 0$ est le risque de base pour une durée t, x un vecteur d'explicatives, et r(x) une fonction de ces explicatives, habituellement un index constitué par une combinaisons linéaire des $\mathbf{x}$.

Le risque devant être positif pour toutes les valeurs des explicatives, on respecte cette exigence en imposant une transformée logarithmique sur l'index :

$$r(\mathbf{x}) = \exp(\mathbf{x}^\top \beta) = exp(x_1\beta_1 + x_2\beta_2 + \ldots) \tag{4.2}$$

Soit au total [4] :

$$h(t) = h_0(t) \exp(\mathbf{x}^\top \beta) \tag{4.3}$$

Notez bien qu'à part la condition $h_0(t) > 0$, aucune hypothèse n'est faite sur le risque de base [5].

Si note δ_i une indicatrice telle que $\delta_i = 1$ si l'événement est observé pour le i$^{\text{ème}}$ individu et $\delta_i = 0$ sinon, alors, dans le cas d'une censure à droite, l'expression de la vraisemblance pour un échantillon constitué de n individus indépendants est :

$$L = \prod_{i=1}^{n} [f_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i} \tag{4.5}$$

Soit $\Re_{t_i}$ les individus à risque au temps d'événement t_i. En supposant pour l'instant qu'il n'y a pas concomitance dans la survenue de l'événement entre plusieurs individus : chaque temps

4. Notez qu'à la différence de l'écriture du modèle linéaire usuel, il n'y a pas de constante dans l'index du modèle de Cox pour la simple raison qu'elle serait indéterminée. En effet, $\forall c, h(t) = [h_0(t) \exp(-c)] \exp(c + \mathbf{x}^\top \beta)$. En conséquence, on force $c = 0$ et le terme constant éventuel est implicitement intégré à la composante *risque de base* $h_0(t)$.

5. Au passage, rappelons que nous avons aussi :

$$S(t; \mathbf{x}, \beta) = [S_0(t)]^{\exp(\mathbf{x}^\top \beta)} \tag{4.4}$$

avec $S_0(t) = \exp[-H_0(t)] = \exp[-\int_0^t h_0(v)dv]$.

d'événement observé est spécifique à un individu et un seul, alors il vient :

$$L = \prod_{i=1}^{n} [h_i(t_i)S_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta_i} \tag{4.6}$$

$$= \prod_{i=1}^{n} [h_i(t_i]^{\delta_i} S_i(t_i) \tag{4.7}$$

$$= \prod_{i=1}^{n} \left[\frac{h_i(t_i)}{\sum_{j\in\Re_{t_i}} h_j(t_i)}\right]^{\delta_i} \left[\sum_{j\in\Re_{t_i}} h_j(t_i)\right]^{\delta_i} S_i(t_i)] \tag{4.8}$$

Cox propose de ne considérer que le premier terme de l'expression précédente pour construire la vraisemblance partielle qui se définit donc comme :

$$PL = \prod_{i=1}^{n} \left[\frac{h_i(t_i)}{\sum_{j\in\Re_{t_i}} h_j(t_i)}\right]^{\delta_i} \tag{4.9}$$

Afin de comprendre la logique qui est en oeuvre ici, il importe de comprendre que le terme

$$\left[\frac{h_i(t_i)}{\sum_{j\in\Re_{t_i}} h_j(t_i)}\right]$$

est une probabilité conditionnelle : c'est la probabilité qu'un individu connaisse l'événement au temps t_i sachant qu'il s'est produit un événement à cette durée parmi tous les individus à risque $\Re_{t_i}$. En effet, si on repart de la définition du risque instantané alors on peut dériver une approximation de la probabilité de survenue de l'événement dans un intervalle de temps Δt contenant t_i pour un individu j à risque en t_i :

$$Pr[t_i \leq T_j < t_i + \Delta t | T_j \geq t_i] = h_j(t_i)\Delta t$$

La probabilité de connaître un événement en t_i étant égale à la probabilité qu'il se produise pour le premier individu à risque en recensé dans $\Re_{t_i}$ ou pour le deuxième ou pour le troisième ou ...., et les survenues de l'événement étudié étant indépendantes entre les individus, on a évidemment

$$Pr[\text{survenue d'un événement en } t_i] = \sum_{j\in\Re_{t_i}} h_j(t_i)\Delta t$$

La probabilité que cet événement concerne l'individu i, celui qui a effectivement connu l'événement étudié, étant $h_i(t_i)\Delta t$, les termes qui définissent la vraisemblance partielle sont donc bien les probabilités conditionnelles annoncées [6] :

$$\frac{h_i(t_i)}{\sum_{j\in\Re_{t_i}} h_j(t_i)} = Pr[i|t_i]$$

Par ailleurs, en reprenant l'équation de base des modèles PH, il vient :

6. Par exemple, avec 3 individus à risque A, B et C ayant des probabilités de connaître l'événement à une durée donnée égales respectivement à P(A), P(B) et P(C). Si les individus sont indépendants, la probabilité de survenue de l'événement a cette durée est donc P(A)+P(B)+P(C). Par ailleurs, P(A)=P(A|A ou B ou C)×P(A ou B ou C) et donc P(A|A ou B ou C)=P(A)/(P(A)+P(B)+P(C).

$$Pr[i|t_i] = \frac{h_i(t_i)}{\sum_{j\in\Re_{t_i}} h_j(t_i)} \tag{4.10}$$

$$= \frac{h_0(t_i)r(\mathbf{x_i})}{\sum_{j\in\Re_{t_i}} h_0(t_i)r(\mathbf{x_j})} \tag{4.11}$$

$$= \frac{\exp\left(\mathbf{x_i}^\top\beta\right)}{\sum_{j\in\Re_{t_i}} \exp\left(\mathbf{x_j}^\top\beta\right)} \tag{4.12}$$

On note avec la dernière égalité que la probabilité qu'un individu connaisse l'événement à un temps t_i ne dépend plus de la durée elle-même, mais seulement de l'ordre d'arrivée des événements. En conséquence La valeur de la vraisemblance partielle est invariante à une transformation monotone des durées, ce qui peut permettre éventuellement d'avancer un argument de robustesse en faveur du modèle de Cox [7].

Cette vraisemblance partielle s'écrit donc :

$$PL = \prod_{i=1}^{n}\left[\frac{\exp\left(\mathbf{x_i}^\top\beta\right)}{\sum_{j\in\Re_{t_i}} \exp\left(\mathbf{x_j}^\top\beta\right)}\right]^{\delta_i} \tag{4.13}$$

L'avantage de recourir à PL est évidemment que cela élimine la référence au risque de base $h_0(t)$ et que donc il n'est plus nécessaire de parier sur une distribution des temps de survie comme il fallait le faire avec l'approche paramétrique. Cet aspect a évidemment fortement contribué à la popularité de cette modélisation.

Il existe cependant un coût : en comparant les expressions de la vraisemblance (4.8) et de la vraisemblance partielle (4.9), on peut noter que les coefficients β sont présents dans les termes omis par PL : les estimateurs obtenus par maximisation de la vraisemblance L ne sont pas identiques à ceux obtenus par maximisation de la vraisemblance partielle PL, ces derniers devant être moins efficaces puisqu'ils sont obtenus en faisant une impasse sur de l'information pertinente. Pour autant, on peut montrer qu'ils vérifient deux propriétés des estimateurs du maximum de vraisemblance : ils sont consistants et asymptotiquement gaussiens.

Afin d'illustrer les développements précédents on peut suivre un exemple de calcul de d'une vraisemblance partielle Supposons que l'on ait un échantillon composé de cinq individus (A, B, C, D, E) avec les temps d'événements respectifs $(5, 8^*, 2, 3^*, 9)$, le signe $*$ signalant une durée censurée à droite. On observe donc 3 durées pour lesquelles un événement s'est réalisé :

— au temps $t_1 = 2$, l'événement ayant concerné l'individu C alors que l'ensemble des individus risqués était $\Re_{t_1} = \{A, B, C, D, E\}$. Le terme associé à ce temps t_1 dans PL sera donc :

$$\frac{r(\mathbf{x}_C)}{r(\mathbf{x}_A) + r(\mathbf{x}_B) + r(\mathbf{x}_C) + r(\mathbf{x}_D) + r(\mathbf{x}_E)}$$

— au temps $t_2 = 5$, l'événement concernant l'individu A alors que l'ensemble des individus risqués à cette durée est $\Re_{t_2} = \{A, B, E\}$. Le terme associé à ce temps t_2 dans PL sera donc :

$$\frac{r(\mathbf{x}_A)}{r(\mathbf{x}_A) + r(\mathbf{x}_B) + r(\mathbf{x}_E)}$$

7. ces temps étant seulement nécessaires pour identifier les individus à risque à chaque date de survenue de l'événement

individu	durée	sexe
1	1	0
2	3	0
3	5	1

TABLE 4.1 – Exemple de données illustrant un cas de monotonicité de la vraisemblance partielle

— Enfin, un dernier événement se produit au temps $t_3 = 9$ pour l'individu E alors qu'il n'était plus que la dernière personne encore risquée et donc $\Re_{t_3} = \{E\}$. Le terme associé à ce temps t_3 dans PL sera donc :

$$\frac{r(\mathbf{x}_E)}{r(\mathbf{x}_E)}$$

Au final, sur cet échantillon, la vraisemblance partielle aura ainsi l'expression suivante :

$$PL = \frac{r(\mathbf{x}_C)}{r(\mathbf{x}_A) + r(\mathbf{x}_B) + r(\mathbf{x}_C) + r(\mathbf{x}_D) + r(\mathbf{x}_E)} \times \frac{r(\mathbf{x}_A)}{r(\mathbf{x}_A) + r(\mathbf{x}_B) + r(\mathbf{x}_E)}$$

Les estimateurs des β sont obtenus en maximisant la log-vraisemblance partielle au moyen d'un algorithme de Newton-Raphson. Il est conseillé de suivre le détail des itérations via l'option `itprint` de la commande `model`. Lorsqu'une divergence des estimateurs est observée, il peut alors être utile de mettre en oeuvre la correction de Firth.

4.1.2 La correction de Firth en cas de monotonicité de PL

Il s'agit d'une configuration que vous pouvez rencontrer en pratique. Elle se manifeste par des valeurs absolues de l'un où de plusieurs des coefficients β estimés extrêmement grandes. Son origine vient de ce qu'une où une combinaison des variables explicatives permet une séparation complète de la variable expliquée. Nous allons prendre un exemple simple : une seule variable explicative, `sexe`, codée 0 pour les hommes et 1 pour les femmes , 3 individus dans l'échantillon, aucune censure et les données présentées dans la table(4.1). Dans ce cas, nous avons un seul coefficient β à estimer, et, avec trois temps d'événements distincts, la vraisemblance partielle à maximiser s'écrit :

$$\begin{aligned} PL &= \left(\frac{e^{\beta\times 0}}{e^{\beta\times 0} + e^{\beta\times 0} + e^{\beta\times 1}}\right) \times \left(\frac{e^{\beta\times 0}}{e^{\beta\times 0} + e^{\beta\times 1}}\right) \times \left(\frac{e^{\beta\times 1}}{e^{\beta\times 1}}\right) \\ &= \left(\frac{1}{2 + e^{\beta}}\right) \times \left(\frac{1}{1 + e^{\beta}}\right) = \frac{1}{2 + 3e^{\beta} + e^{2\beta}} \end{aligned}$$

et

$$\frac{\delta PL}{\delta \beta} = -\frac{3e^{\beta} + 2e^{2\beta}}{(2 + 3e^{\beta} + e^{2\beta})^2} < 0$$

En conséquence, pour maximiser PL, $\hat{\beta}$ va tendre en valeur absolue vers l'infini. A titre d'exercice, vous pouvez vérifier que si nous avions utilisé le codage opposé, à savoir, sexe=0 pour les hommes et 1 pour les femmes, alors la dérivée de la vraisemblance partielle obtenue serait positive pour toute valeur de β, i.e. pour la maximisation, $\hat{\beta}$ tendrait vers $+\infty$. Dans cet exemple, la séparation parfaite tient à ce que les durées longues sont exclusivement le fait des femmes.

Dans ce cas, Firth a proposé de maximiser une vraisemblance partielle qui incorpore un terme de pénalité reposant sur la matrice d'information de Fisher estimée. La log-vraisemblance pénalisée est donnée par

$$log(PL^*) = log(PL) + \frac{1}{2}log(|I(\beta)|) \tag{4.14}$$

avec $I(\beta) = -\frac{\delta log(PL)}{\delta \beta^2}$. On peut montrer que $log(PL^*)$ est concave en β même lorsque $log(PL)$ est monotone. En conséquence, les valeurs qui maximisent $log(PL^*)$ restent finies même en cas de parfaite séparation [8]. Certains auteurs, comme Allison, préconisent même l'emploi systématique de la correction de Firth : les estimateurs obtenus sur petits échantillons par maximisation de $log(PL^*)$ auraient notamment un biais sur petits échantillons inférieur à celui afférent à ceux obtenus via la maximisation de $log(PL)$. En prolongement de cette dernière remarque, notez qu'il est conseillé d'utiliser cette correction de Firth lorsqu'on est en présence d'événements rares, *i.e.* lorsque le nombre d'événement observés constitue un échantillon de faible taille [9].

En pratique dans PHREG, il suffit d'ajouter l'option `firth` à la commande `model`. N'hésitez pas à lire l'exemple consacré à la correction de Firth donné dans l'aide de la proc PHREG. Retenez cependant que l'emploi de cette option invalide les intervalles de confiance construits avec les statistiques de Wald sur les ratios de risque au moyen des commandes `contrast` et `hazardratio` discutées un peu plus loin.

4.1.3 La prise en compte d'événements simultanés

En principe, les durées étant supposées être les réalisations d'une aléatoire continue, la probabilité que deux individus ou plus connaissent l'événement étudié au même instant est nulle. En pratique cependant les données sont souvent relevées selon une certaine discrétisation du temps (données hebdomadaires, mensuelles, etc...) de sorte qu'il est courant qu'une même durée de survie soit afférente à plusieurs individus. Fondamentalement cela ne complique pas sérieusement les écritures précédentes, simplement le nombre de calculs peut devenir tel si on utilise la contribution exacte de ces individus à la vraisemblance partielle que des approximations ont été proposées dont trois sont implémentées sous PHREG. Le plus simple est de présenter ces ajustements au moyen d'un exemple. Supposons que pour une certaine durée deux individus A et B aient connu l'événement et que trois autres C, D, E soient à risque. Pour simplifier les écritures, on notera $r_i = \exp(\beta^\top \mathbf{x}_i), i = A, B, C, D, E$.

- TIES=EXACT : On considère que l'égalité des durées pour A et B n'est pas réelle mais est associée uniquement à l'imprécision de la collecte des données. Dans la réalité, A doit avoir connu l'événement avant B ou B doit l'avoir connu avant A. Avec l'option EXACT, tous les cas possibles sont considérés. Ainsi, si on considère que A connaît l'événement avant B, alors lorsque l'on traite le cas de A, il y a 5 individus à risque et lorsqu'on traitera de B, il n'y aura plus que 4 individus risqués puisque A est censé avoir disparu. La probabilité

8. Le choix de ce terme de pénalité est fondé par une argumentation d'économétrie bayésienne que l'on ne peut pas expliciter ici.

9. Remarquez encore qu'une analyse dite d'événements rares est dénommée ainsi en raison de l'effectif des événements observés et non pas de leur proportion dans l'échantillon total.

d'observer les deux événements simultanément sera donc :

$$\left(\frac{r_A}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_B}{r_B + r_C + r_D + r_E}\right) \tag{4.15}$$

mais il est tout aussi plausible que B ait connu l'événement avant A et dans ce cas cette probabilité serait :

$$\left(\frac{r_B}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_A}{r_A + r_C + r_D + r_E}\right) \tag{4.16}$$

Les deux classements étant possibles, la contribution de la durée en question à la vraisemblance partielle est finalement égale à :

$$\begin{aligned}\left(\frac{r_A}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_B}{r_B + r_C + r_D + r_E}\right) \\ + \left(\frac{r_B}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_A}{r_A + r_C + r_D + r_E}\right)\end{aligned} \tag{4.17}$$

Si 3 individus A, B et C ont la même durée d'événement pour 5 individus risqués, nous aurions eu six termes :

$$\begin{aligned}&\left(\frac{r_A}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_B}{r_B + r_C + r_D + r_E}\right)\left(\frac{r_C}{r_C + r_D + r_E}\right) \\ &+ \left(\frac{r_A}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_C}{r_B + r_C + r_D + r_E}\right)\left(\frac{r_B}{r_B + r_D + r_E}\right) \\ &+ \left(\frac{r_B}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_A}{r_A + r_C + r_D + r_E}\right)\left(\frac{r_C}{r_B + r_D + r_E}\right) \\ &+ \left(\frac{r_B}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_C}{r_A + r_C + r_D + r_E}\right)\left(\frac{r_A}{r_A + r_D + r_E}\right) \\ &+ \left(\frac{r_C}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_A}{r_A + r_B + r_D + r_E}\right)\left(\frac{r_B}{r_B + r_D + r_E}\right) \\ &+ \left(\frac{r_C}{r_A + r_B + r_C + r_D + r_E}\right)\left(\frac{r_B}{r_A + r_B + r_D + r_E}\right)\left(\frac{r_A}{r_A + r_D + r_E}\right)\end{aligned} \tag{4.18}$$

Plus généralement, si k individus ont le même temps d'événement, alors la vraisemblance partielle pour ce temps sera composé de $k!$ termes. Dans ces conditions, le temps de calcul peut devenir pénalisant. Par exemple, avec k=10 on aura 3.628.800 termes à calculer. Dans des conditions réelles d'études, le nombre d'événements simultanés peut devenir beaucoup plus important, si bien que l'on conçoit aisément la nécessité de recourir à des approximations de ce calcul exact

- TIES=BRESLOW. La vraisemblance partielle est approximée par

$$PL_{\texttt{BRESLOW}} = \prod_{i=1}^{k} \frac{\prod_{j \in \mathfrak{D}_{t_i}} r_j}{\left[\sum_{j \in \Re(t_i)} r_j\right]^{d_i}} \tag{4.19}$$

où où $\mathfrak{D}_{t_i} = \{i_1, i_2, \ldots, i_{d_i}\}$ sont les indices des d_i individus pour qui l'événement se réalise en t_i. Cette approximation évite notamment d'avoir à réévaluer les dénominateurs des différents termes de la vraisemblance partielle. Par exemple, si on reconsidère le cas des 5 individus précédents, avec 2 événements simultanés, on aura, à la place de (4.17) l'évaluation de

$$\frac{r_A r_B}{(r_A + r_B + r_C + r_D + r_E)^2}$$

Du fait de la simplicité des calculs, cette méthode est une des plus populaires. C'est d'ailleurs cette approximation qui est mise en oeuvre par défaut dans la proc PHREG. La littérature souligne cependant qu'elle conduit à des estimations des coefficients β qui peuvent être fortement biaisés vers zéro notamment lorsque le nombre d'événements concomitants d_i est élevé relativement à l'effectif à risque en t_i.

- TIES=EFRON. Il s'agit ici est de corriger le dénominateur utilisé dans l'approximation de Breslow. Si on compare celui-ci à celui du calcul exact, on peut s'apercevoir qu'il est à l'évidence trop élevé. L'idée d'Efron est d'introduire la moyenne des risques des individus ayant connu l'événement et non pas leur niveau de risque total. Formellement l'expression de la log-vraisemblance devient :

$$PL_{\texttt{EFRON}} = \prod_{i=1}^{k} \frac{\prod_{j \in \mathfrak{D}_{t_i}} r_j}{\prod_{j=1}^{d_i} \left[\sum_{l \in \mathfrak{R}_{t_i}} r_l - \frac{j-1}{d_i} \sum_{l \in \mathfrak{D}_{t_i}} r_l \right]} \tag{4.20}$$

Avec le même exemple que précédemment, on aura ainsi à la place de (4.17)

$$\frac{r_A r_B}{(r_A + r_B + r_C + r_D + r_E)(0.5 r_A + 0.5 r_B + r_C + r_D + r_E)}$$

La littérature sur le sujet s'accorde pour conseiller l'utilisation de l'approximation d'Efron à celle de Breslow : même si son temps d'exécution est plus élevé, elle semble fournir des estimations des coefficients β toujours plus proches de ceux obtenus avec la méthode exacte.

— TIES=DISCRETE. Cette option se démarque fondamentalement des trois précédentes du fait que le temps est ici considéré comme étant discret : si des événement concomitants sont observés, c'est qu'ils se sont réellement produits simultanément et il n'y a plus aucune raison de chercher à imaginer des classements dans leur ordre d'arrivée.
Rappelons qu'au temps t_i nous avons d_i individus, dont les indices sont repérés par $\mathfrak{D}_{t_i} = \{i_1, i_2, \ldots, i_{d_i}\}$, qui ont connu l'événement étudié alors qu'il y avait $\mathfrak{R}_{t_i}$ individus risqués. La probabilité que se soit précisément ces individus référencés par $\mathfrak{D}_{t_i}$ qui aient connu l'événement est donc la probabilité de prendre d_i individus parmi tous les individus à risque, soit :

$$\frac{\prod_{j \in \mathfrak{D}_{t_i}} r_j}{\sum_{P_i} \prod_{j \in P_i} r_j} \tag{4.21}$$

Le dénominateur étant la somme sur toutes les façons possibles de prendre d_i éléments parmi tous ceux à risque en t_i. Dans l'exemple de nos 5 individus avec deux événements concomitants, ces deux événement aurait pu être
(A,B),(A,C),(A,D),(A,E),
(B,C),(B,D),(B,E),
(C,D),(C,E),
(D,E).

Pour évaluer (4.21) il faudrait calculer la somme des probabilités de chacun de ces couples, puis diviser la probabilité associé à l'événement observé, soit (A,B) par cette somme. La vraisemblance partielle dans ce cas, originellement également proposée par Cox devient alors :

$$PL_{\texttt{DISCRETE}} = \prod_{i=1}^{k} \frac{\prod_{j \in \mathfrak{D}_{t_i}} r_j}{\sum_{P_i} \prod_{j \in P_i} r_j} \tag{4.22}$$

La difficulté avec cette option est relative au temps de calcul nécessaire pour évaluer le dénominateur lorsque d_i est élevé.

Pour résumer, le choix entre les diverses options précédentes doit d'abord se faire en fonction d'une réponse à la question suivante : le temps est-il vraiment discret, les événements concomitants se réalisent-ils effectivement au même instant ? ou est-il continu, les événements apparemment concomitants se produisant en fait à des durées différentes mais cachées notamment en raison du mode de collecte des données ? Une réponse positive à la première question doit logiquement entraîner le choix de l'option TIES=DISCRETE, sachant que les temps de calculs peuvent alors devenir rédhibitoires. Une réponse négative à cette même question menant à l'une des trois autres options. Entre celles-ci, le choix de TIES=EXACT s'impose si le nombre d'événements simultanés n'est pas trop élevé, si le temps de calcul devient excessif, et si le nombre d'individus à risque est faible par rapport aux nombre d'individus à risque, alors BRESLOW et EFRON vont donner pratiquement des résultats identiques. Lorsque $\frac{card(\mathfrak{D}_{t_i})}{card(\mathfrak{R}_{t_i})}$ devient relativement élevé alors l'option TIES=EFRON doit être préférée malgré un temps de calcul plus important que BRESLOW.
Pour clore cette section, signalons que même si le temps est discret, l'option TIES=DISCRETE n'est pas obligatoirement la panacée : nous verrons dans le chapitre suivant qu'en utilisant les recommandations de Brown, de Allison et de Shumway comment on peut estimer un modèle de durée à temps discret en passant par la proc logistique et remplacer avantageusement la vraisemblance partielle de la proc PHREG et son option TIES=DISCRETE par une vraisemblance complète.

4.1.4 Spécification de l'équation à estimer, commandes `Model` et `Class`

La commande `Model` est obligatoirement présente dans un appel de `PHREG`. C'est par elle que l'on va indiquer quels sont les temps d'événements à expliquer, s'il s'agit de temps censurés ou non, et la liste des variables explicatives. La commande `Class` est facultative, on sait qu'elle est particulièrement utile lorsque l'on désire estimer l'impact de variables catégorielles. Cependant, son implémentation dans la proc PHREG permet une meilleur maîtrise de la catégorie qui est mise en référence notamment par rapport à ce qui est fait dans la proc LIFEREG où la base est toujours la dernière catégorie, celle-ci étant définie par l'option `ORDER.`

La commande `Model`

On va trouver deux formes possibles pour la commande `Model`, l'une traditionnelle, identique à celle déjà vue dans la proc `LIFEREG`, l'autre, plus récemment implémentée dans SAS relève d'une approche en termes de processus de comptage est sur certains aspects beaucoup plus souple que la précédente. Le choix de l'une ou l'autre forme n'est pas anodin puisque la structure des données en est directement affectée. Par exemple, avec la première nous aurons en principe un seul

enregistrement par individu présent dans l'échantillon de travail, alors que la seconde réclamera généralement plusieurs enregistrements par individu.

1. La première version. On précise la nom de la variable contenant les temps d'événements, éventuellement le signe * suivi du nom d'une autre variable et une liste de valeurs de censure entre parenthèses : si pour un individu l'observation de la deuxième variable appartient à cette liste, alors, pour cet individu le temps indiqué est un temps d'événement censuré à droite. Si elle n'y appartient pas, alors l'événement s'est réalisé au temps indiqué. Suit ensuite le signe = puis la liste des variable explicatives des durées. Ainsi :

   ```
   model time = x1 x2;
   model time*cens(0,1,3) = x1 x2;
   ```

   Dans le premier exemple il n'y a pas de censure. Dans le second, si $cens_i \in \{0, 1, 3\}$ alors $time_i$ correspond à une durée censurée à droite, sinon, l'événement s'est effectivement réalisé en $time_i$ pour l'individu de rang i.

2. La seconde version est adaptée à une structure de données de type *processus de comptage* : chaque enregistrement précise les bornes t_1 et t_2 d'un intervalle de durées de la forme $]t_1, t_2]$, ouvert à gauche et fermé à droite c'est à dire tel qu'un événement qui se réalise en t_1 n'appartient pas à cet intervalle alors qu'il y appartient s'il survient en t_2. La variable de censure a le même comportement que dans la version précédente mais n'est relative qu'à la durée t_2. Dans ce format, un individu donné est souvent décrit par plusieurs enregistrements s'il est observé sur plusieurs sous-périodes. Soit par exemple, un client masculin (variable genre codée 1) qui a été suivi pendant 24 mois après un premier achat. Il a repassé une commande 8 mois après, puis une autre 20 mois après cet achat initial. A la fin des 24 mois, il n'a plus repassé de commande. Cette personne sera décrite par trois enregistrements :

b_1	b_2	cens	genre
0	8	1	1
8	20	1	1
20	24	0	1

   La commande sera alors de la forme

   `model (`t_1`, `t_2`)*cens(0) = genre ...;`

   Un avantage important de ce second style est la facilité de prise en compte d'explicatives non constantes dans le temps. Cette possibilité sera examinée plus en détail par la suite, mais pour illustration, supposons que dans la base de données il y ait une variable contact (codée 1, si le client a été visité, 0 sinon). Si le client précédent a reçu un représentant pendant les mois 0 à 8, puis pendant les mois 20 à 24, on aurait les données du tableau suivant, avec

b_1	b_2	cens	genre	contact
0	8	1	1	1
8	20	1	1	0
20	24	0	1	1

   la commande associée :

`model (`t_1`, `t_2`)*cens(0) = genre contact ...;`

On notera enfin que si la borne basse d'un intervalle de durées, t_1 est souvent la borne haute de l'intervalle immédiatement précédent, ce n'est pas obligatoire : les intervalles peuvent être discontinus.

La commande `Class`

Appliquée à une variable catégorielle, elle va créer automatiquement un ensemble d'indicatrices qui pourront être intégrées dans l'équation à estimer. Pour l'essentiel, elle évite simplement le passage par une étape `Data` dans laquelle on créerait les indicatrices en question. A priori son emploi simplifie la construction de l'équation à estimer, limite les risques d'erreur et facilite la maintenance du programme puisqu'un recodage de la variable catégorielle initiale est automatiquement pris en compte, sans que l'on ait à reprendre les codes de l'étape `Data`.

Le point essentiel est évidemment de connaître les modalités de création des indicatrices et quelles valeurs leurs sont attribuées pour pouvoir interpréter correctement les résultats des ajustements. Par ailleurs, même dans le cas simple où une indicatrice binaire, 0/1, est créée, il faut bien évidemment savoir ce que signifient le "1" et le "0". Le design des indicatrices est géré par l'option `PARAM`. Nous allons examiner seulement les deux valeurs les plus courantes données à cette option en supposant l'existence de deux variables :

— Groupe, avec les modalités "A", "B"et "C",
— Genre, binaire 0/1 avec le format : 1 →"Femmes",0 →"Hommes"

1. PARAM=GLM. Soit la commande

   ```
   class groupe genre / param=glm;
   ```

   Dans ce cas, la Class Level Information table, affichée dans l'output est celle de la table 4.2. Elle précise que 3 indicatrices ont été créée pour la variable Groupe, la première valant 1 pour toutes les personnes appartenant à la catégorie A, 0 sinon, la seconde égale à 1 pour toutes les persones de la catégorie B, 0 sinon et la troisième valant 1 pour les personnes de la catégorie C, 0 sinon. Deux indicatrices obéissant au même principe ont également été créées pour la variable Genre, la première valant 1 si la personne est une femme, 0 si s'est un homme, et la seconde égale à 0 si la personne est une femme, 1 si s'agit d'un homme. Rappelez-vous que par défaut l'option `ORDER=FORMATTED` est activée Si à la suite de cette commande `Class` on utilise comme explicatives les variables Group et Genre, alors la dernière catégorie de chacune est utilisée comme base, soit ici la catégorie "C" et le genre "Hommes". On retrouve exactement le comportement de l'instruction `Class` de la proc LIFEREG. En particulier, pour changer la catégorie de référence, il faut recoder les variables Groupe et Genre, ou leur appliquer un nouveau format, de sorte que `ORDER=FORMATTED` classe en dernière catégorie celle que l'on désire mettre en base. Une autre possibilité est d'utiliser l'option `ref=first`, et dans ce cas la première catégorie de chaque variable est mise en base [10]. Ainsi, avec les données de notre exemple, mettra en base les femmes de la catégorie "A" avec

   ```
   class groupe genre / param=glm ref=first;
   ```

10. On peut aussi utiliser `param=glm ref=last;` qui ne fait que reproduire le comportement par défaut de `param=glm;`.

Class	Value	Design		
Groupe	A	1	0	0
	B	0	1	0
	C	0	0	1
Genre	Femmes	1	0	
	Hommes	0	1	

TABLE 4.2 – Class Level Information table associée à PARAM=GLM

Class	Value	Design	
Groupe	A	1	0
	B	0	0
	C	0	1
Genre	Femmes	1	
	Hommes	0	

TABLE 4.3 – Class Level Information table associée à PARAM=REF

Dans le tableau présentant les résultats de l'estimation, on repère la catégorie de référence du fait que son coefficient est exactement égal à zéro, sans écart-type estimé.

2. PARAM=REF. Du fait de son ancienneté, le codage `Param=glm` est souvent connu. On peut cependant penser que les avantages du codage `param=ref`, notamment pour désigner la catégorie de référence, vont faire qu'elle va devenir assez rapidement la paramétrisation la plus utilisée. C'est d'ailleurs la valeur par défaut de l'option `param=`.
On vient de voir que dans le cas d'une variable catégorielle comprenant c modalités, l'option `param=glm` créait c variables indicatrices, puis PHREG forçait à zéro le coefficient de l'indicatrice de référence. L'option `param=ref` va créer seulement $c-1$ indicatrices en ne créant pas celle qui est mise en référence explicitement par l'utilisateur [11]. Pour ce faire, dans la commande `Class` on doit naturellement lister les variables catégorielles à considérer, mais aussi la modalité de référence pour chacune d'elle. Ainsi les deux commandes :

```
class groupe(ref='B') genre(ref='Hommes') / param=ref;
class groupe(ref='B') genre(ref='Hommes');
```

sont équivalentes [12] et conduisent à la Class Level Information table 4.3

11. Cette modalité de référence n'apparaîtra donc pas dans la table des résultats de l'ajustement.

12. équivalentes puisque `param=ref` est utilisé par défaut. Habituez-vous cependant à faire apparaître explicitement la valeur des options, cela facilite la compréhension lors des relectures ultérieures du programme que l'on vient d'écrire, et limite les sources d'erreur notamment lorsque les valeurs par défaut changent lorsqu'on passe d'une procédure à l'autre. Par exemple, dans SAS 9.4, comme on l'a signalé, l'instruction `Class` de la proc LIFEREG a comme défaut param=glm, alors que cette même instruction a pour défaut `param=ref` dans la proc PHREG. Rappelons que dans la proc LOGISTIC, `Class` a comme défaut param=effect, où les indicatrices sont codées -1 dans la modalité de référence. Ainsi, dans notre exemple, `class groupe/param=effect` créerait 2 indicatrices selon la Class Level Information table suivante :

Class	Value	Design	
Groupe	A	1	0
	B	0	1
	C	-1	-1

4.2 Les ratios de risque

4.2.1 Interprétation des coefficients et Ratios de Risque

Nous avons déjà vu l'écriture de base du modèle de COX pour un individu quelconque décrit par ses caractéristiques $x_1, x_2, \ldots, x_k$, :

$$\begin{aligned} h_(t) &= h_0(t) r(\mathbf{x}) \\ &= h_0(t) \exp(\beta^\top \mathbf{x}) \\ &= h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \end{aligned} \tag{4.23}$$

— cas d'une explicative x_j continue. Il vient :

$$\begin{aligned} dh(t) &= h_0(t) dr(\mathbf{x}) \\ &= h_0(t) \frac{\delta r(\mathbf{x})}{\delta x_j} dx_j \\ &= \beta_j [h_0(t) r(\mathbf{x})] dx_j \end{aligned} \tag{4.24}$$

Le terme entre crochets étant positif, le signe de l'impact d'une explicative sur le risque est le même que le signe du coefficient de cette explicative. Le lien étant non linéaire, l'ampleur de l'impact ne doit cependant pas être confondue avec la valeur de ce coefficient.

— cas d'une explicative discrète : on effectue directement le calcul avec les valeurs des modalités. Soit x_j une variable prenant deux valeurs a ou b , alors l'écart de risque entre deux individus ne se distinguant que par cette valeur de x_j est donné par :

$$\begin{aligned} \Delta h(t) &= h_0(t) \Delta r(\mathbf{x}) \\ &= h_0(t) [r(\mathbf{x}) | x_j = b] - r(\mathbf{x}) | x_j = a] \end{aligned} \tag{4.25}$$

La présence de $h_0(t)$ fait que l'on préfère mesurer l'impact d'une variable sur le risque en termes relatifs. En effet, pour deux individus, l et m identiques en tous points, sauf au regard de la $j^{\text{ième}}$ explicative, le ratio de risque est donné par :

$$\begin{aligned} RR &= \frac{h_l(\mathbf{x}) | x_j = a)}{h_m(\mathbf{x}) | x_j = b)} \\ &= \frac{e^{\beta_j a}}{e^{\beta_j b}} \\ &= e^{\beta_j (a-b)} \end{aligned} \tag{4.26}$$

Si les deux valeurs x_{lj} et x_{mj} sont séparée d'une unité, on a évidemment

$$RR = e^{\beta_j} \tag{4.27}$$

soit encore, toujours pour une modification unitaire de l'explicative, une variation relative du risque égale à $e^{\beta_j} - 1$. Par exemple, si $\beta_j = 0.5$, alors le ratio de risque vaut 0.60, ce qui signifie indifféremment

— qu'une augmentation d'un point de l'explicative x_j provoque une baisse de 40% (=0.60-1) du risque, toutes autres variables explicatives inchangées,
— où que le risque d'un individu pour lequel l'explicative en question augmente d'un point représente 60% du risque qu'il supportait avant l'augmentation.

On conçoit donc aisément que le rendu des résultats d'une estimation d'un modèle de Cox fasse plus souvent référence aux ratios de risque, immédiatement intelligibles, qu'aux coefficients des explicatives eux-mêmes. Lorsque le ratio de risque peut se calculer sans ambiguïté, PHREG affiche sa valeur et on dispose dans la commande `model` de l'option `risklimits` faisant apparaître les bornes d'un intervalle de Wald sur le ratio dans le tableau des résultats de l'estimation. Pour le calcul des ratios dans les cas plus complexes, deux instructions sont disponibles depuis SAS 9.2 : `Hazardratio` et `Contrast`. La dernière est la plus générale : tout ce que fait `Hazardratio` peut être également réalisé avec `Contrast`, l'inverse n'étant pas vrai. La première est en revanche censée être plus simple d'utilisation,

4.2.2 Commandes Hazardratio et Contrast

Vous avez compris qu'un ratio de risque est une mesure du risque supporté par un individu relativement à un autre. A l'issue de l'ajustement d'un modèle de Cox ayant k explicatives, le risque estimé pour deux individus quelconque i et j est respectivement de :

$$\hat{h}_{t_i} = \hat{h}_0(t)\exp(\hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}),$$
$$\hat{h}_{t_j} = \hat{h}_0(t)\exp(\hat{\beta}_1 x_{j1} + \hat{\beta}_2 x_{j2} + \ldots + \hat{\beta}_k x_{jk}),$$

et l'évaluation du ratio de risque de i relativement à j est simplement :

$$\widehat{RR}_{i/j} = exp(\hat{\beta}_1[x_{i1} - x_{j1}] + \hat{\beta}_2[x_{i2} - x_{j2}] + \ldots + \hat{\beta}_k[x_{ik} - x_{jk}]) \tag{4.28}$$

ou encore, en notant x_i et x_j les caractéristiques des individus i et j, et $\hat{\beta}$ le vecteur des coefficients estimés, *i.e* :

$$x_i = (x_{i1}, x_{i2} + \ldots, x_{ik})^\top$$
$$x_j = (x_{j1}, x_{j2} + \ldots, x_{jk})^\top$$
$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots + \hat{\beta}_k)^\top$$

$$\widehat{RR}_{i/j} = exp(\hat{\beta}^\top[x_i - x_j]] \tag{4.29}$$

Son calcul est donc particulièrement simple et le plus difficile est sans doute à chercher dans la syntaxe des instructions

L'instruction Contrast

Avec cette instruction, on devra indiquer la valeur du coefficient de pondération de chacun des coefficients $\hat{\beta}$ estimés. Pour cela, il faut naturellement savoir identifier l'explicative correspondante à chaque coefficient. Cela est évidemment simple pour les variables continues dans la mesure ou le nom de la variable identifie également son coefficient. En revanche, en présence de variables catégorielles, le nom de la variable va souvent renvoyer à plusieurs indicatrices et donc plusieurs

coefficients. Dans ce cas, l'instruction `Contrast` exige que l'on ait parfaitement compris le fonctionnement des options `PARAM=` et `REF=` de la commande `CLASS`. Dans ce qui suit, nous utiliserons toujours le système d'indicatrices créées par l'option `PARAM=REF`.

Afin d'illustrer la mise en oeuvre de `Contrast`, on va supposer à nouveau que nous cherchions à expliquer le risque au moyen des explicatives suivantes :

— Une variable "Groupe" ayant trois modalités, ou catégories, "A", "B" et "C",
— Une variable "Genre", binaire 0/1 sur laquelle on a appliqué le format : 1 →"F",0 →"H",
— Une variable continue "Age"

Supposons l'exécution des deux commandes ci-dessous. Notez l'introduction d'interactions entre Groupe et Genre.

```
class group(ref='B') genre(ref='H') / param=ref;
model time*cens(0) = age groupe genre groupe*genre;
```

L'équation ajustée est donc :

$$\log(h_t) = \log(h_{0t}) + \beta_{age}AGE + \beta_a * (\text{Group=A}) + \beta_c * (\text{Group=C}) + \beta_f * (\text{Genre=F}) + \beta_{af} * (\text{Group=A \& Genre=F}) + \beta_{cf} * (\text{Group=C \& Genre=F})] \quad (4.30)$$

avec entre parenthèse la valeur d'une indicatrice valant 1 si la variable est égale à la modalité précisée, 0 sinon. On a bien deux indicatrices associées à la variable Groupe, avec la catégorie "B" en référence, une indicatrice associée à Genre, avec les hommes en référence, et les deux indicatrices associées aux interactions. Soit, dans l'ordre de sortie des résultats :

— Un coefficient pour la variable "Age",
— Deux coefficients pour la variable "GROUP", le premier afférent à l'indicatrice (Groupe=A), le second afférent à l'indicatrice (Groupe=C)
— Un coefficient pour la variable "Genre" afférent à l'indicatrice (Genre=F)
— Deux coefficient pour la variable d'interaction "Groupe*Genre", le premier associé aux femmes du groupe A, le second aux femmes du groupe C.

En d'autres termes, aux variables "Age" ou "Genre" devront être associé un seul coefficient de pondération. A la variable "Groupe" il faudra préciser la valeur de deux pondérations, et se souvenir qu'elles s'appliquent dans l'ordre au coefficient de l'indicatrice (Groupe=A) puis au coefficient de (Groupe=C). Enfin, il faudra aussi préciser deux pondérations pour la variable "Genre*Groupe", la première touchant le coefficient de l'indicatrice (Groupe=A & Genre=F), la seconde celui de l'indicatrice (Groupe=C & Genre=F).
Construisons maintenant des exemples d'appel de `Contrast`.

1. Comment évolue le risque avec l'âge ? Si l'individu i est identique à l'individu j en tout point à l'exception de l'âge : il est un an plus vieux. On a évidemment, d'après (4.28) ou (4.29)

   $$RR_{i/j} = e^{\hat{\beta}_{age}}$$

   L'instruction `Contrast` associée est la suivante :

   ```
   Contrast "RR, 1 an de plus" age 1 group 0 0 genre 0 groupe*genre 0 0 / estimate=exp;
   ```

   que l'on pourra simplifier en

```
Contrast "RR, 1 an de plus" age 1 / estimate=exp;
```

On peut en effet ne pas faire apparaître les variables pour lesquelles tous les coefficients associés sont affectés d'une pondération nulle.
L'option `estimate=exp` demande à `Contrast` de travailler sur l'exponentielle de la combinaison linéaire des $\hat{\beta}$. En son absence, elle afficherait le résultat de la combinaison linéaire elle-même, c'est à dire [13] $log(RR_{i/j}) = \hat{\beta}^\top [x_i - x_j]$, mais on sait que l'on préfère généralement raisonner sur les ratios de risque, plutôt que sur les coefficients eux-mêmes.

Assez souvent pour les variables continues, on veut apprécier l'évolution du risque pour des variations supérieures à l'unité, par exemple, quel est le ratio de risque entre deux personnes identiques à l'exception de l'âge, i étant plus âgé de 5 ans que j ? Dans ce cas, toujours selon (4.28)

$$RR_{i/j} = e^{\hat{\beta}_{age}[Age_i - Age_j]} = e^{5\hat{\beta}_{age}} = \left(e^{\hat{\beta}_{age}}\right)^5$$

et la commande `Contrast` correspondante est :

```
Contrast "RR, 5 ans de plus" age 5 / estimate=exp;
```

2. On veut comparer le risque afférent à des hommes identiques à l'exception de leur groupe d'appartenance A,B ou C. Les équations de définition des ratios de risque qui nous intéressent sont alors :

$$\begin{aligned}
\mathrm{RR}_{\text{A versus B, genre=H}} &= \exp[\beta_a], \\
\mathrm{RR}_{\text{A versus C, genre=H}} &= \exp[\beta_a - \beta_c] \\
\mathrm{RR}_{\text{C versus B, genre=H}} &= \exp[\beta_c]
\end{aligned}$$

et les instructions `contrast` correspondantes :

```
Contrast 'A versus B, Hommes' group 1 0 / estimate=exp;
Contrast 'A versus C, Hommes' group 1 -1 / estimate=exp;
Contrast 'C versus B, Hommes' group 0 1 / estimate=exp;
```

Notez que l'on peut, dans le fichier de sortie, faire afficher un texte entre quotes. Généralement, cette étiquette rappelle la signification du test affiché.

3. On veut comparer le risque afférent à des femmes identiques entre elles à l'exception de leur groupe d'appartenance A,B ou C. Il vient :

$$\begin{aligned}
\mathrm{RR}_{\text{A versus B, genre=F}} &= \exp[\beta_a + \beta_{af}] \\
\mathrm{RR}_{\text{A versus C, genre=F}} &= \exp[\beta_a + \beta_{af} - \beta_c - \beta_{cf}] \\
\mathrm{RR}_{\text{C versus B, genre=F}} &= \exp[\beta_c + \beta_{cf}]
\end{aligned}$$

et les instructions `contrast` correspondantes :

```
Contrast 'A versus B, Femmes' group 1 0 Group*genre 1 0 / estimate=exp;
Contrast 'A versus C, Femmes' group 1 -1 Group*genre 1 -1 / estimate=exp;
Contrast 'C versus B, Femmes' group 0 1 Group*genre 0 1 / estimate=exp;
```

13. Cf. (4.29.

En plus du ratio de risque, `Contrast` donne un intervalle de confiance [14] à $100(1-\alpha)\%$, où α est un seuil de risque géré par l'option `Alpha=` .

On peut aussi tester l''égalité des risques des individus i et j, soit H0 :$RR_{i/j} = 1$ versus $RR_{i/j} \neq 1$, via des tests LRT, Wald, Lagrange [15]. Ainsi, les trois précédentes instructions `Contrast` vont faire apparaître, entre autres résultats, trois tests de Wald, chacun à 1 degré de liberté. Le dernier, par exemple, aura comme hypothèse nulle, l'égalité des risques de femmes du même âge, les unes appartenant au groupe C, les autres au groupe B. Notez enfin qu'il serait possible de tester l'hypothèse jointe d'égalité des risques de femmes du même âge indépendamment de leur groupe, soit, pour un âge (quelconque) donné :

HO : risque des femmes du groupe A = risque des femmes du groupe B
&
risque des femmes du groupe A = risque des femmes du groupe C
&
risque des femmes du groupe C = risque des femmes du groupe B

Il suffirait pour cela de regrouper les trois commandes `Contrast` précédentes en une seule, en séparant leurs arguments par une virgule. Soit :

```
Contrast 'A versus B, A versus C, C versus B, Genre=Femmes'
group 1 0 Group*genre 1 0 ,
group 1 -1 Group*genre 1 -1 ,
group 0 1 Group*genre 0 1 / estimate=exp;
```

La statistique, sous H0, aurait une distribution de χ^2 à 2 degrés de liberté [16].

L'instruction Hazardratio

A priori plus simple à utiliser que l'instruction `Contrast` dans la mesure où elle ne suppose pas la connaissance de la structure et des valeurs des indicatrices créées par `Class` : elle réclame le nom de la variable pour laquelle on veut calculer le ou les ratios de risque, et les valeurs des autres variables qui vont conditionner le calcul

1. **Cas d'une variable continue** : il s'agit du cas le plus simple on a vu que pour une modification d'une unité, le ratio de risque associé est simplement e^{β}, où β est le coefficient de la variable. Dans la table de sortie standard affichant les résultats de l'estimation d'un modèle, la colonne intitulée `"Hazard Ratio"` donne déjà les estimations de ceux-ci précisément pour une augmentation d'une unité de l'explicative concernée. Les valeurs affichées [17]

14. Construit à partir du Chi2 de Wald. On rappelle que cet intervalle est invalidé par l'utilisation de l'option `firth` dans la commande `model`.

15. La syntaxe est la suivante :

`Contrast ... / alpha=`α `test= mot-clef;`

où mot-clef $\in$ {NONE, ALL, LR, WALD, SCORE} et $|\alpha| < 1$. Par défaut `test=wald`. Les bornes de l'intervalle de confiance sont également calculées avec une statistique de Wald.

16. Pourquoi 2 degrés de liberté ?

17. Lorsqu'il s'agit du coefficient d'une variable dont l'impact dépend de la valeur d'une autre explicative, typiquement en cas d'interactions explicitées dans la commande `Model`, PHREG ne calcule pas, et n'affiche donc pas, de ratio de risque pour cette variable.

sont donc simplement les exponentielles des coefficients $\hat{\beta}$ indiqués en première colonne du même tableau. Vous pouvez cependant être amenés à calculer un rapport de risque pour des augmentations non unitaires . Par exemple si l'âge est une explicative, il peut être plus intéressant de calculer le ratio de risque entre deux personnes qui diffèrent de 5 ou 10 années, plutôt que d'une seule. Dans ce cas, il suffit d'exécuter la commande `hazardratio` en précisant dans l'option `units=` la valeur de la variation dont vous désirez apprécier l'impact. Vous avez également la possibilité de réclamer la construction d'un intervalle de confiance via l'option `cl=` à un seuil $(1-\alpha)$, et de préciser la valeur de votre seuil de risque α dont la valeur par défaut est comme toujours, de 5%. Par exemple :

```
hazardratio age / units=5 alpha=0.10 cl=wald;
hazardratio age / units=10 cl=pl;
hazardratio '5 ans de plus' age / units=5 cl=both;
```

Vous constatez dans ces exemples que PHREG peut fournir deux estimations différentes pour les bornes de l'intervalle de confiance [18].

(a) La première, que vous connaissez, est réclamée par le mot clef `wald`. Elle repose sur la propriété de normalité asymptotique des estimateurs des coefficients $\hat{\beta}$ estimés par maximisation d'une vraisemblance, et sur la propriété d'invariance de ces estimateurs qui assure qu'asymptotiquement $e^{\hat{\beta}}$ est également gaussien, puisqu'estimateur du maximum de vraisemblance de e^{β}. Il suffit donc de calculer l'écart-type $s_{e^{\hat{\beta}}}$ du rapport de risque connaissant l'écart-type de $\hat{\beta}$ via la méthode delta pour finalement obtenir l'intervalle cherché comme $e^{\hat{\beta}} \pm q_{1-\alpha/2} \times s_{e^{\hat{\beta}}}$, où $q_{1-\alpha/2}$ est la quantile d'ordre $1-\alpha/2$ de la gaussienne standardisée.

(b) La seconde, associée au mot clef `pl`, pour `"Profile Likelihood"` est fondé sur un travail de Venzon et Moolgavkar et est censée fournir des estimateurs des bornes d'intervalle plus robustes que les précédentes, notamment lorsqu'on travaille avec de petits échantillons et que l'on doute de la normalité des estimateurs [19]. Leur solution repose sur l'approximation asymptotique du test LRT par un χ^2, le gain provenant du fait que le test LRT approcherait sa distribution asymptotique plus rapidement que le test de Wald. Notez encore que des travaux semblent montrer qu'en cas d'événements rares, au moins pour la régression logistique [20], l'emploi de la correction de Firth et des estimateurs PL seraient préférables aux estimateurs standards.

Dans le dernier des trois exemples précédents on demande l'affichage des deux types d'estimations des bornes d'un intervalle de confiance à 95%.

2. **Cas d'une variable catégorielle** Il s'agit alors d'apprécier l'évolution du risque lorsque deux individus sont semblables en tout point à l'exception de leur catégorie d'appartenance. Dans ce cas, l'instruction demande que l'on précise l'option `dif=` avec deux valeurs possibles, `dif=all` qui demande le calcul des ratios de risque de toutes les catégories de la variables

18. Encore une fois, l'emploi de l'option `firth` dans la commande `model` conduit à des bornes d'intervalle du type Wald erronées.

19. les estimations des bornes des intervalles `"Profile-likelihood"` sont obtenues au moyen d'un processus itératif. En cas de difficulté, il est possible d'intervenir sur ce processus via des options `PLCONV=`, `PLMAXIT=`, `PLSINGULAR=`. Voir l'aide de PHREG pour plus de détails à ce sujet.

20. Par exemple, G. Heinze et M. Schemper, A solution to the problem of separation in logistic regression, *Statist. Med.*, pp 2409-2419, 2002.

prises deux à deux, ou `dif=ref` qui évalue les ratios de risque de chacune des catégories relativement à la seule catégorie de référence. Si ces ratios dépendent des modalités où valeurs d'autres variables, il suffit de préciser avec l'option `at` les modalités où les valeurs pour lesquelles on désire faire le calcul du ou des ratios de risque. Supposons toujours les commandes :

```
class group(ref='B') genre(ref='H') / param=ref;
model time*cens(0) = age group genre groupe*genre;
```

Si nous désirons comparer pour des hommes du même age, le risque qu'il y a a appartenir aux divers groupes A, B et C, il suffira de faire :

```
hazardratio 'Hommes, compare A,B,C' Group / diff=all at (genre='H');
```

La comparaison des risques de femmes appartenant aux groupes A et C à celui des femmes de même âge de la catégorie B, qui est en base, sera obtenue avec :

```
hazardratio 'Femmes, compare A et C à B' Group / diff=ref at (genre='F');
```

Pour comparer les risques des hommes et des femmes de la catégorie A par exemple, on fera :

```
hazardratio 'Hommes versus Femmes, Groupe A' Genre / diff=ref at (groupe='A');
```

3. cas d'une interaction continue*catégorielle. Supposons la suite de commandes qui introduit une telle interaction :

```
class group(ref='B') genre(ref='H') / param=ref;
model time*cens(0) = age group genre age*genre;
```

L'équation estimée est alors :

$$\begin{aligned}\log(h_t) = \log(h_{0t}) + \beta_{age} AGE + \beta_a * (\text{Group=A}) + \beta_c * (\text{Group=C}) + \beta_f * (\text{Genre=F}) \\ + \beta_{a,age} * [\text{Age} \times (\text{Genre=F})]\end{aligned} \tag{4.31}$$

Supposons que l'on veuille comparer les risques entre hommes et femmes âgés de 25 ans. La commande `Contrast` serait :

```
Contrast'Femmes vs Hommes, 25 ans' genre 1 age*genre 25;
```

L'instruction `Hazardratio` équivalente étant [21] :

```
Hazardratio 'Femmes vs Hommes, 25 ans' genre / diff=ref at (age=25);
```

4.2.3 Des exemples de sorties

Afin d'illustrer quelques uns des développements qui précèdent, nous allons reprendre les données contenues dans le fichier uis_small déjà analysées au moyen de la procédure LIFEREG. L'équation alors ajustée est celle correspondante à la commande model ci-après, qui est naturellement maintenant exécutée au sein de la proc PHREG. Toujours à des fins d'illustration, nous avons également intégré des formats sur les variables *treat* et *site*. Soit :

21. Pour encore plus d'exemples, voyez Paul T. Savarese et Michael J. Patetta, An Overview of the Class, Contrast and Hazardratio Statements in the SAS 9.2 PHREG Procedure, Paper 253-2010, SAS Global Forum 2010.

Analysis of Maximum Likelihood Estimates										
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
age		1	-0.03369	0.00929	13.1512	0.0003	.	.	.	Age at Enrollment
ndrugtx		1	0.03646	0.00770	22.4092	<.0001	1.037	1.022	1.053	Number of Prior Drug Treatments
treat	long	1	-0.26741	0.09123	8.5921	0.0034	0.765	0.640	0.915	Treatment Randomization Assignment long
site	B	1	-1.24593	0.50873	5.9979	0.0143	.	.	.	Treatment Site B
age*site	B	1	0.03377	0.01551	4.7423	0.0294	.	.	.	Treatment Site B * Age at Enrollment

TABLE 4.4 – PHREG : Exemples de sortie - Tableau des paramètres estimés

```
proc format;
value prog 1="long" 0="court";
value lieu 1="B" 0="A";
run;
proc phreg data=uis;
format treat prog.;
format site lieu.;
class treat(ref="court") site(ref="A") / param=ref;
model time*censor(0) = age ndrugtx treat site age*site / risklimits;
contrast "RR, 20 ans, site B versus A" site 1 age*site 20 / estimate=exp;
contrast "RR, 30 ans, site B versus A" site 1 age*site 30 / estimate=exp;
contrast "RR, + 5 ans, site A" age 5 / estimate=exp;
contrast "RR, +5 ans, site B" age 5 age*site 5 / estimate=exp;
hazardratio "RR, 20 ans, site B versus A" site / diff=ref at (age=20) cl=both;
hazardratio "RR, 30 ans, site B versus A" site / diff=ref at (age=30) cl=both;
hazardratio "RR, 5ans de plus, site A" age / units=5 at (site="A") cl=both;
hazardratio "RR, 5ans de plus, site B" age / units=5 at (site="B") cl=both;
run; quit;
```

Les résultats de la commande `model` concernant les coefficients estimés sont présentés dans la table 4.4. Je vous laisse interpréter ces résultats, et retrouver les valeurs affichées pour les ratios de risque. Notez qu'en raison de l'interaction de l'âge et du site, PHREG n'affiche pas ce ratio pour ces deux variables, ni pour le produit croisé. La raison en est bien évidemment qu'avec cette spécification, le risque associé à l'âge dépend du site et réciproquement : il est nécessaire de spécifier les valeurs de ces deux variables pour calculer des ratios de risque. C'est précisément ce que l'on fait avec les commandes contrast et hazardratio. Notez qu'en deuxième colonne, PHREG indique la modalité dont le coefficient est estimée. Il s'agit par exemple du site "B", ce qui était évidemment attendu puisque nous avons spécifié que le site "A" devait être mis en base [22].

Les résultats associés aux 4 instructions `contrast` sont présentés dans la table 4.5. La première indique que le risque de rechute pour un individu de 20 ans hospitalisé dans le centre "B" représente 56% du risque d'une personne de 20 ans hospitalisé dans le site "A". A 30 ans, la seconde instruction estime que ce risque est remonté à 80%. La troisième évalue le risque de rechute d'une personne qui, hospitalisé dans le site A, aurait 5 ans de plus qu'une autre personne ayant les mêmes caractéristiques et également hospitalisée en "A" : le risque de la première est estimé à 80%

22. En l'absence de la proc format, c'est le programme long et le site B qui auraient été mis en base (codage=1, versus codage 0 pour le programme court, et le site A).

Contrast Estimation and Testing Results by Row									
Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi-Square	Pr > ChiSq
RR, 20 ans, site B versus A	EXP	1	0.5653	0.1206	0.05	0.3720	0.8588	7.1449	0.0075
RR, 30 ans, site B versus A	EXP	1	0.7924	0.0834	0.05	0.6447	0.9738	4.8924	0.0270
RR, + 5 ans, site A	EXP	1	0.8450	0.0393	0.05	0.7714	0.9255	13.1512	0.0003
RR, +5 ans, site B	EXP	1	1.0004	0.0626	0.05	0.8849	1.1310	0.0000	0.9950

TABLE 4.5 – PHREG : Exemples de sortie - Résultats associés aux instructions `contrast`

RR, 20 ans, site B versus A: Hazard Ratios for site					
Description	Point Estimate	95% Wald Confidence Limits		95% Profile Likelihood Confidence Limits	
site B vs A At age=20	0.565	0.372	0.859	0.370	0.855

RR, 30 ans, site B versus A: Hazard Ratios for site					
Description	Point Estimate	95% Wald Confidence Limits		95% Profile Likelihood Confidence Limits	
site B vs A At age=30	0.792	0.645	0.974	0.643	0.971

RR, 5ans de plus, site A: Hazard Ratios for age					
Description	Point Estimate	95% Wald Confidence Limits		95% Profile Likelihood Confidence Limits	
age Unit=5 At site=A	0.845	0.771	0.926	0.771	0.925

RR, 5ans de plus, site B: Hazard Ratios for age					
Description	Point Estimate	95% Wald Confidence Limits		95% Profile Likelihood Confidence Limits	
age Unit=5 At site=B	1.000	0.885	1.131	0.884	1.130

TABLE 4.6 – PHREG : Exemples de sortie - Résultats associés aux instructions `hazardratio`

du risque de la seconde. La dernière instruction `contrast` refait les mêmes calculs mais pour deux personnes hospitalisées dans le site "B". Dans ce cas, deux personnes séparées par 5 années supporterait le même niveau de risque [23]. Les tests de Chi2 associés indiquent que les ratios de risque sont significativement différents de l'unité aux seuils usuels à l'exception du quatrième [24].

Dans la table 4.6 on trouve les résultats des instructions `hazardratio`. Celles-ci correspondent dans l'ordre aux mêmes interrogations que celles considérées via les commandes `contrast` précédemment discutées [25]. Il est donc logique que les estimateurs des ratios de risque soient identiques dans les deux tables. La seule différence concerne l'apparition des bornes des intervalles de confiance PL. On constate d'ailleurs qu'elles ne décalent pratiquement pas, dans cet exemple, de celles obtenues avec Wald.

4.3 L'estimation de la survie de base

L'estimation de la survie de base est importante puisque c'est à partir d'elle que l'on va pouvoir estimer les fonctions de survie de tout individu ayant les caractéristiques décrites dans le vecteur des explicatives $\mathbf{x}$. On connaît déjà les relations liant survie, risque instantané et risque cumulé :

$$S(t) = \exp\left\{-\int_0^t h(\tau)d\tau\right\} = \exp\{-H(t)\}$$

23. A titre d'exercice, retrouvez la syntaxe de ces instructions

24. Ce que l'on peut évidemment deviner en notant que la valeur 1 est dans l'IC de la quatrième instruction mais n'appartient pas aux trois premiers IC.

25. Toujours à titre d'exercice, vous devriez pouvoir reconstruire leur syntaxe, et au passage remarquez la plus grande facilité dans la prise en compte de la variable d'interaction : elle doit être explicitement traitée avec `contrast`, elle est automatiquement gérée par `hazardratio`.

et donc, dans le cadre du modèle de Cox :

$$S(t, \mathbf{x}) = \exp\left\{-\int_0^t h_0(\tau)\exp\left[\beta^\top \mathbf{x}\right]d\tau\right\} \tag{4.32}$$

$$= \exp\left\{-\int_0^t h_0(\tau)d\tau\right\}^{\exp\left[\beta^\top \mathbf{x}\right]} \tag{4.33}$$

$$= S_0(t)^{\exp\left[\beta^\top \mathbf{x}\right]} \tag{4.34}$$

où $S_0(t) = \exp\left\{-\int_0^t h_0(\tau)d\tau\right\}$ est la survie de base, *i.e.* la survie afférente à un individu pour lequel toutes les explicatives seraient nulles.

PHREG propose trois estimateurs non paramétriques de cette survie de base :

1. L'estimateur de Breslow :
   C'est une extension de l'estimateur de Nelson-Aalen en présence de variables explicatives. Le nombre attendu d'événements au temps t_i est égal à la somme des risques instantanés afférents aux individus à risque en t_i : $E(d_i) = \sum_{j\in\Re_{t_i}} h_{0_{t_i}} e^{\hat{\beta}^\top \mathbf{x}_j}$. Si on égalise cette valeur attendue à la valeur observée, il vient [26] :

   $$\hat{h}_{0_{t_i}} = \frac{d_i}{\sum_{j\in\Re_{t_i}} e^{\hat{\beta}^\top \mathbf{x}_j}} \tag{4.35}$$

   La fonction de risque de base cumulée estimée est alors donnée par

   $$\hat{H}_{0_t} = \sum_{i|t_i<t} \hat{h}_{0_{t_i}} \tag{4.36}$$

   On utilise enfin l'équation qui relie risque cumulé et survie pour obtenir l'estimateur de la survie de base :

   $$\hat{S}_{0_t} = e^{-\hat{H}_{0_t}} \tag{4.37}$$

   Notez que les estimateurs précédents sont évalués sur les temps d'événements effectifs : $\hat{H}_{0_t}$ et $\hat{S}_{0_t}$ sont des fonctions en escaliers.
2. L'estimateur de Flemming-Harrington :
   Introduit à partir de SAS 9.4, il modifie l'estimateur de Breslow en cas d'événement simultanés. Alors que dans le précédent les individus concernés ont une pondération unitaire, de sorte que le numérateur de 4.35 est égal à d_i, ils vont ici recevoir des pondérations différenciées.

26. Hanley (2008) précise élégamment le lien qui existe entre l'équation (4.35) et l'estimateur usuel de Kaplan-Meier. Dans ce dernier, il n'y a pas d'explicatives, tous les individus sont supposés homogènes, et le risque serait estimé par $d_i/card(\Re_{t_i})$. La prise en compte des ratios de risque au dénominateur revient à retrouver l'équivalent d'un effectif d'individus homogènes à partir d'ensembles d'individus hétérogènes. On reprend son exemple : soit un modèle ayant une seule explicative, sexe=0 si la personne est une femme, 1 sinon, et son coefficient estimé $\hat{\beta} = 0.4054$, avec donc $\exp(\hat{\beta}) = 1.5$. Si à un temps d'événement t_i on a 50 femmes et 60 hommes à risque alors, le dénominateur de (4.35) est égal à $(50 \times 1 + 60 \times 1.5) = 140$. *i.e.* à ce temps, on a l'équivalent de 140 femmes. L'estimateur KM dans cette population homogène constituée uniquement de femmes serait égal à $d_i/140$, et c'est cette quantité qui est recréée par (4.35).

3. L'estimateur Product-Limit, ou de Kalbfleish et Prentice :
   On considère un modèle à temps discret et on note π_{0_i} la probabilité conditionnelle de connaître l'événement à une durée t_i pour un individu en base, *i.e.* un individu pour lequel les explicatives sont toutes de valeur nulle. On sait que la survie d'un individu ayant des caractéristiques $\mathbf{x}_j$ est donnée par $S_j(t) = S_0(t)^{\exp(\beta^\top \mathbf{x}_j)}$ et, si k est la nombre total d'événements observés, la vraisemblance est donnée par [27]

$$L = \prod_{i=1}^{k} \prod_{j \in \mathfrak{D}_i} (1 - \pi_{0_i}^{e^{\beta^\top \mathbf{x}_j}}) \prod_{j \in \mathfrak{R}_i - \mathfrak{D}_i} \pi_{0_i}^{e^{\beta^\top \mathbf{x}_j}} \tag{4.38}$$

   Si on prend comme estimation de β les valeurs obtenues par maximisation de la vraisemblance partielle alors, en l'absence d'événements simultanés, la maximisation de cette vraisemblance par rapport à π_{0_i} a pour solution :

$$\hat{\pi}_{0_i} = \left(1 - \frac{e^{\hat{\beta}\mathbf{x}_i}}{\sum_{j \in \mathfrak{R}_i} e^{(\hat{\beta}^\top \mathbf{x}_j)}}\right)^{e^{-\hat{\beta}^\top \mathbf{x}_i}} \tag{4.39}$$

   En présence d'événements simultanés, il n'y a pas de solution explicite et une méthode itérative doit être mise en oeuvre pour la maximisation de L.
   Pour finir, une fois les π_{0_i} estimés, l'estimateur de la survie de base est égal à :

$$S_{0_t} = \prod_{i | t(i) \leq t} \hat{\pi}_{0_i} \tag{4.40}$$

En pratique, lorsque les effectifs des ensembles d'individus ayant les mêmes temps d'événements ne sont pas trop grands par rapport à ceux des individus à risque, les courbes de survie calculées par ces trois estimateurs sont généralement proches.

Le choix entre l'une où l'autre des procédures d'estimation est précisé au moyen de l'option `method=` dans la commande `BASELINE`, selon :
— `method=BRESLOW` pour la première,
— `method=FH` pour la deuxième, et
— `method=PL` pour l'estimateur Product-Limit
Par défaut, l'estimateur retenu est BRESLOW.

La commande `BASELINE` permet de référencer un fichier contenant les valeurs des explicatives pour lesquelles on désire estimer la fonction de survie. Par défaut, les calculs sont effectués en prenant la moyenne d'échantillon des explicatives continues et les modalités de référence pour les variables catégorielles. En conséquence, comme on sait que la survie de base est obtenue pour des

27. Un individu ayant les caractéristiques $\mathbf{x}_j$ a une probabilité de connaître l'événement après t_i sachant qu'il ne l'a pas encore connu avant t_i donnée par :

$$\frac{S(t_i)}{S(t_{i-1})} = \frac{S_0(t_i)^{\exp(\beta^\top \mathbf{x}_j)}}{S_0(t_{i-1})^{\exp(\beta^\top \mathbf{x}_j)}} = \pi_{0i}^{\exp(\beta^\top \mathbf{x}_j)}$$

et donc une probabilité de réalisation de l'événement en t_i égale à $1 - \pi_{0i}^{\exp(\beta^\top \mathbf{x}_j)}$

valeurs nulles des explicatives, il faut éviter ce comportement par défaut et donc créer effectivement le fichier utilisé en input par `BASELINE`. Si on reprend l'exemple précédent, l'obtention de la survie de base S_{0t} et du risque cumulé de base s'effectuera au moyen des commandes suivantes :

création du fichier contenant les caractéristiques des individus pour lesquels on veut la survie estimée :

```
data base0;
input age ndrugtx treat site;
format treat prog.;
format site lieu.;
cards;
0 0 0 0
0 0 0 0
;
run;
```

appel de `phreg` avec `model=...` pour l'estimation de notre équation. La commande `baseline` prend ensuite les données de la table précédente, grâce à `covariate= base0`, et créé la table `survie0`, qui contiendra, pour chaque temps d'événements effectif, les estimations de la survie de base, variable 'Sbreslow' et du risque cumulé, variable 'Hbreslow' calculées avec l'estimateur de Breslow :

```
proc phreg data=uis;
format treat prog. site lieu.;
class treat(ref="court") site(ref="A") / param=ref;
model time*censor(0) = age ndrugtx treat site;
baseline out=survie0 covariates=base0 survival=Sbreslow cumhaz=Hbreslow /
method=breslow;
run;
quit;
```

En répétant ces calculs avec les options `method=pl` et `method=fh`, on obtient les survies et les risques cumulés de base estimés avec les deux autres procédures. Les graphes de ces fonctions sont fournis dans la figure 4.1. Dans cet exemple, et comme cela est généralement le cas, les valeurs estimées par ces trois méthodes sont proches et conduisent à des courbes qui se superposent presque parfaitement.

4.4 L'analyse stratifiée avec le modèle de Cox

On se rappelle que le modèle de Cox est un modèle à risque proportionnel : le ratio de risque entre deux individus i et j dont les caractéristiques sont données respectivement par x_i et x_j est

$$RR_{i/j} = \frac{h_{it}}{h_{jt}} = \frac{h_{0t}e^{(\beta^\top x_i)}}{h_{0t}e^{(\beta^\top x_j)}} = e^{(\beta^\top [x_i - x_j])}$$

Si les explicatives sont invariantes dans le temps, alors ce ratio est une constante relativement aux durées : si le ratio de risque entre i et j est égal à 3 pour une durée de 3 mois, il doit être égal à 3

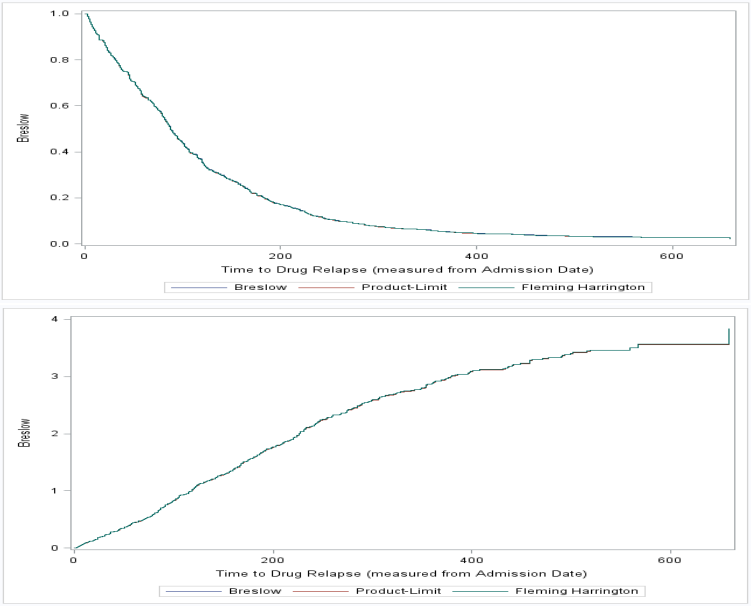


FIGURE 4.1 – Graphes de la survie et du risque cumulés de base obtenus avec les 3 procédures d'estimation

pour une durée de 12 mois, de 24 mois, etc...Il peut arriver que cette hypothèse paraisse erronée : en général, cela survient notamment lorsque l'on a une hétérogénéité au sein de la population étudiée, par exemple une différenciation de la clientèle selon le sexe, le statut familial, le type de contrat, etc...On peut par exemple imaginer que le ratio de risque entre hommes et femmes soit à une certaine valeur pour des durées courtes, puis que les comportements des individus diffèrent de sorte que pour des durées moyennes et élevées ce ratio se modifie. En matière de contrat, on peut supposer que la fidélisation à moyen/long terme associée à un type de contrat ne soit pas la même pour un autre type : si l'événement étudié est le non renouvellement du contrat, alors le ratio de risque calculé sur deux clients ne possédant pas le même contrat peut se modifier avec les durées, invalidant l'hypothèse PH.
Dans ces cas de figure, on perçoit l'intérêt d'une analyse stratifiée : elle va autoriser la modification du ratio de risque entre les individus appartenant à des strates différentes. L'abandon de l'hypothèse PH est cependant partiel, puisqu'elle continuera d'être valide pour les membres d'une même strate. Imaginons ainsi que le variable de stratification soit le sexe. L'analyse stratifiée va autoriser une modification du ratio de risque entre hommes et femmes selon les durées, mais elle continuera à d'imposer la constance du ratio de risque entre deux femmes ainsi que sa constance entre deux hommes, et qui plus est, au même niveau dans les deux cas : si le ratio est égal à 3 entre deux femmes ayant certaines caractéristiques, il sera aussi de 3 entre deux hommes ayant les mêmes caractéristiques et cela sur toutes les durées possibles. Pour obtenir ce résultat, l'analyse stratifiée va imposer des coefficients identiques sur les explicatives mais va estimer des risques de base spécifiques à chaque strate. Ainsi, dans le cas de deux strates A et B, nous avons :

$$h_{t_i} = \begin{cases} h_{0,A}(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \quad \texttt{si} \quad i \in A, \\ h_{0,B}(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \quad \texttt{si} \quad i \in B. \end{cases}$$

Dans ces conditions, si deux individus i et j sont dans la même strate, alors :

$$RR_{i/j}(t) = e^{(\beta^\top [x_i - x_j])}$$

et nous savons déja que ce ratio est une constante si les explicatives sont elles-mêmes invariantes dans le temps : l'hypothèse PH est bien vérifiée. En revanche, si $i \in A$ et $j \in B$ alors

$$RR_{i/j}(t) = \frac{h_{0,A}(t)}{h_{0,B}(t)} e^{(\beta^\top [x_i - x_j])} \tag{4.41}$$

Les risques de base pouvant évoluer différemment selon les durées, le terme $h_{0,A}(t)/h_{0,A}(t)$, et en conséquence $RR_{i/j}$, ne sont sont plus tenus d'être des constantes sur l'échelle des durées.

Techniquement, une fonction de vraisemblance partielle est construite sur chaque strate avec le même jeu de coefficients pour les explicatives. Les estimateurs $\hat{\beta}$ sont obtenus par maximisation du produit de ces vraisemblances partielles, et des estimations non paramétriques du risque de base sont ensuite réalisées au sein de chacune des strates.
Un inconvénient de cette levée partielle de l'hypothèse PH est qu'on ne dispose pas avec cette méthode d'une mesure quantifiée de l'impact de la variable de stratification sur le risque : une variable dont les modalités définissent les strates ne peut en effet pas apparaître dans la liste des explicatives, puisque c'est une constante dans les divers sous-ensembles qu'elle définit : son impact est transféré dans la composante 'risque de base'.

Un intérêt de cette méthodologie est qu'elle peut être utilisée pour tirer une information visuelle sur le respect de l'hypothèse PH, i.e., on peut réaliser des ajustements stratifiés pour éventuellement révéler qu'il n'était pas nécessaire de stratifier. L'idée est la suivante : sous l'hypothèse PH les risques sont proportionnels pour deux individus ayant des caractéristiques constantes dans le temps même s'ils n'appartiennent pas à la même strate. Ainsi, si Z est la variable de stratification possédant deux modalités "A" et "B", et pour deux individus i et j tels que $i \in A$ et $j \in B$, alors sous hypothèse de risque proportionnel :

$$h_i(t) = \lambda h_j(t)$$

soit encore $S_i(t) = S_j(t)^\lambda$ et donc $\log(-\log(S_i(t)) = \log(\lambda) + \log(-\log(S_j(t))$. Au final, sous l'hypothèse PH, les courbes de survie estimées pour les individus i et j devraient être parallèles entre elles dans un graphique ayant les durées en abscisse et leurs transformations $log - log$ en ordonnée. Il suffit donc de créer deux individus i et j appartenant à des strates différentes, de demander à PHREG de fournir une estimation de leurs courbes de survie, d'opérer une transformation simple et de finir par un graphique. On illustre la procédure au moyen du fichier 'uis' utilisé dans une sous-section précédente. Comme variable de stratification nous allons utiliser la variable de site, à deux modalités 0 et 1, sur laquelle opère un format (0 →"A", et 1 →"B"). Le point de départ consiste donc en la création d'un fichier de données, contenant nos deux individus. Arbitrairement, nous allons prendre deux personnes de 40 ans (age=40), ayant déjà subis quatre traitements dans le passé (ndrugtx=4) et soumis dans le cadre de l'évaluation actuelle à un traitement court (treat=0). La première est affecté au site "A", la seconde au site "B". Préalablement à la création de ce fichier nous rappelons les formats utilisés.

```
proc format;
value prog 1="long" 0="court";
```

```
value lieu 1="B" 0="A";
run;
data verifPH;
input age ndrugtx treat site
format treat prog.;
format site lieu.;
cards;
40 4 0 0
40 4 0 1
;
run;
```

L'étape suivante consiste à demander l'estimation du modèle via la commande `model`. Par rapport aux estimations qui précèdent, nous avons donc retiré toute référence à la variable `site` dans la liste des explicatives puisqu'il s'agit de la variable de stratification. Il suffit ensuite de réclamer l'estimation de la survie des individus dont les valeurs des explicatives sont dans le fichier `verifPH`. C'est ce que va faire la commande `BASELINE` du programme ci-dessous. Celle-ci prend en entrée le contenu de verifPH (option `covariate`=nom du fichier contenant les valeurs désirées pour les explicatives [28, 29]), et créé en sortie le fichier surverif (option `out=nom de fichier`), qui contiendra les estimateurs réclamés ainsi que les valeurs des explicatives du fichier référencé par covariate, de sorte à pouvoir aisément retrouver dans la nouvelle table à quel individu se rapporte telle ou telle estimation. Dans le cas présent le mot clef `loglogs` demande la sauvegarde, sous le nom 'lls' de la transformée $\log - \log(\hat{S}_t)$. L'option `rowid=site` dans `baseline` permettra d'affecter les observations créées, et donc les courbes affichées par `sgplot`, à chacun des sites [30].

```
proc phreg data=uis;
format treat prog.;
format site lieu.;
class treat(ref="court") / param=ref;
model time*censor(0) = age ndrugtx treat;
strata site;
baseline out=surverif covariates=verifPH loglogs=lls /rowid=site;
run;
quit;
```

On obtient alors les estimations de la table 4.7. On observera que les estimations des coefficients des explicatives restantes ne sont pratiquement pas affectées par la levée de la contrainte de l'hypothèse PH sur la variable de `site` (voir la table 4.5). Pour finir, je vous laisse juge de décider du parallélisme des deux courbes du graphe 4.2, et donc de la validité ou non de cette hypothèse.

```
proc sort data=surverif;
```

28. Notez que dans la commande `baseline` si vous ne spécifiez pas de fichier via `covariate`=, PHREG va prendre par défaut un individu qui aura comme caractéristiques pour les explicatives numériques la valeur moyenne de chaque variable dans la strate, et la modalité de référence pour chaque variable catégorielle.

29. Notez également que `baseline` peut naturellement s'employer sans `strata` : on peut par exemple demander des estimations de la survie, où du risque cumulé pour des individus types à partir d'une estimation non stratifiée.

30. Pour une raison que j'ignore, il faut intercaler une proc sort avant l'appel de `sgplot`. En son absence, celle-ci relie le premier et le dernier point des deux courbes.

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
age		1	-0.02178	0.00751	8.4137	0.0037	0.978	0.964	0.993	Age at Enrollment
ndrugtx		1	0.03522	0.00767	21.0837	<.0001	1.036	1.020	1.052	Number of Prior Drug Treatments
treat	long	1	-0.24445	0.09058	7.2831	0.0070	0.783	0.656	0.935	Treatment Randomization Assignment long

TABLE 4.7 – Paramètres estimés avec la commande `strata site`

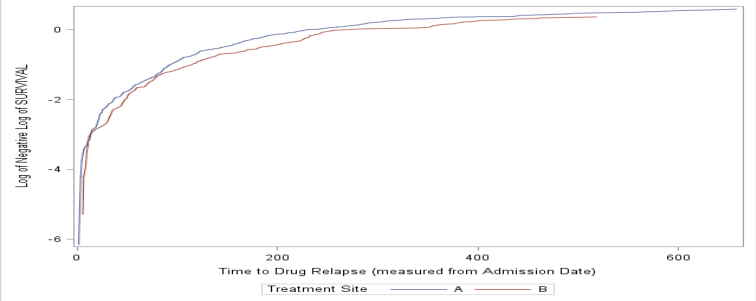


FIGURE 4.2 – log $-log(S[t|$`site="A"`$])$ versus log $-log(S[t|$`site="B"`$])$

```
by site time;
run;
proc sgplot data=surverif;
series y=lls x=time / group=site ;
run;
```

4.5 Explicatives non constantes dans le temps

Une possibilité appréciable du modèle de Cox est d'autoriser des variables explicatives dont les valeurs se modifient avec les durées étudiées. Techniquement, cette facilité vient du fait que la construction de la vraisemblance partielle se fait sur chaque temps d'événement effectif et que pour un temps donné, seuls les individus à risque sont considérés. En conséquence, si on est en mesure de construire, pour chaque individu à risque et pour chaque durée effective, les valeurs des explicatives, il n'y a pas de difficulté particulière à la prise en compte d'explicatives non constantes.

On peut illustrer la motivation à introduire de telles variables au moyen d'un exemple simple. Supposons qu'on étudie un événement pour lequel on soupçonne que le risque de survenue est lié au statut familial des personnes concernées. Soit la variable binaire "statut_ini" codée 1 pour les personnes vivant en couple lors de leur entrée dans l'échantillon et 0 sinon. L'échantillon initial est donc constitué de célibataires et d'individus mariés. Au fur et à mesure que le temps s'écoule, des événements d'intérêt vont se réaliser, définissant les durées effectives sur lesquelles la vraisemblance partielle sera estimée, mais parallèlement, des personnes célibataires peuvent se marier, d'autres peuvent divorcer et donc le statut familial d'un individu est susceptible de se modifier avec les durées d'événement imposant d'ajuster en conséquence le contenu de la variable "couple". Supposons que l'on dispose des variables "duree_evenement", contenant les durées de survenue de l'événement étudié, "duree_statut", contenant la durée de survenue d'un

id	statut _ini	duree _evenement	duree _statut
1	0	8	.
2	1	12	10
3	0	9	8
4	1	7	5

TABLE 4.8 – Exemple de variables explicatives non constantes

changement de statut, codée valeur manquante s'il n'y a pas de changement de statut sur la fenêtre d'observation. Ces deux variables étant naturellement mesurées sur la même échelle des temps. Toujours pour simplifier, on suppose qu'aucun temps d'événement n'est censuré [31] et un seul changement de statut familial sur la période d'observation. Les autres explicatives $x_1, \ldots, x_k$ sont invariantes dans les temps.
La table 4.8 donne quelques exemples d'individus qui pourraient être dans l'échantillon. de travail.

Le premier est entré en tant que célibataire, il a connu l'événement après 8 mois et est resté célibataire sur toute la durée de l'étude. Le quatrième était marié au début de l'étude, a divorcé après 5 mois et a connu l'événement après sept mois.

Deux possibilités sont offertes pour prendre en compte ces explicatives non constantes dans PHREG : l'une adopte en entrée une structure de données de type processus de comptage, l'autre créée les variables explicatives par programme au sein de la procédure.

4.5.1 Données entrées selon un processus de comptage

Comme discuté lors de la présentation de la commande `model`, on aura au besoin plusieurs enregistrements par individus et une variable de censure, ici 'cens', codée 1 si l'événement étudié c'est réalisé, et 0 sinon. L'organisation des observations pour nos quatre individus est décrite dans la table 4.9. Dans cette table, la variable 'statut' est l'explicative décrivant le statut familial sur chaque intervalle de durée. La commande model correspondante sera de la forme :

```
model (t1, t2)*cens(0) = statut x1 ...xk;
```

4.5.2 Explicatives non constantes créées par programme

Nous aurons dans ce cas un seul enregistrement par individu avec des observations telles que présentées dans la table 4.8.
Nous indiquons dans la table 4.10 les valeurs que doit prendre la variable 'statut' pour construire la vraisemblance partielle à chaque temps d'événement effectif et pour chacune des personnes repérées par leur identifiant.

L'appel de PHREG adapté pourrait être le suivant. [32] :

31. Dans le cas contraire, il suffirait d'introduire une indicatrice de censure sur "duree_evenement"
32. Si l'événement et le changement de statut surviennent à la même durée, on va considèrer arbitrairement que le changement de statut précède l'événement.

id	t1	t2	cens	statut	x_1	...	x_k
1	0	8	1	0	x_{11}	...	x_{11}
2	0	10	0	1	x_{12}	...	x_{12}
2	10	12	1	0	x_{12}	...	x_{12}
3	0	8	0	0	x_{13}	...	x_{13}
3	8	9	1	1	x_{13}	...	x_{13}
4	0	5	0	1	x_{14}	...	x_{14}
4	5	7	1	0	x_{14}	...	x_{14}

TABLE 4.9 – Explicatives non constantes, structure des données en entrée de type processus de comptage

duree _evenement	id=1	id=2	id=3	id=4
7	0	1	0	0
8	0	1	1	/
9	/	1	1	/
12	/	0	/	/

TABLE 4.10 – Valeurs attendues pour "statut" pour chaque individu selon les durées d'événement. La case est marquée / si l'individu n'est plus à risque.

```
proc phreg;
model duree_evenement= statut x1 x2 ... xk;
if duree_statut>duree_evenement or missing(duree_statut)=1
then statut=statut_ini;
else statut=1-statut_ini;
run;
```

4.6 Tests de validation

Valider un modèle est un exercice sans fin du fait du nombre de directions à regarder. Quelques aspects doivent cependant être systématiquement examinés :

— La qualité de l'ajustement,
— La spécification retenue
— Le test de l'hypothèse PH
— La détection des outliers et des observations influentes

Plusieurs résidus se révèlent alors utiles pour l'examen de ces questions : les résidus de martingale pour l'étude de la spécification de l'équation, les résidus de déviance pour la recherche des observations mal expliquées par le modèle correspondant à des outliers, les résidus du score pour celle des observations influentes et les résidus de Schoenfeld pour juger de la validité de l'hypothèse PH. Avant de présenter ces divers résidus, nous commençons par rappeler des tests déjà connus permettant de juger de la qualité de l'ajustement au regard de la pertinence des explicatives retenues.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.7306	8	<.0001
Score	34.7775	8	<.0001
Wald	34.4983	8	<.0001

TABLE 4.11 – Tests de nullité de l'ensemble des coefficients

4.6.1 La qualité de l'ajustement

La pertinence peut être appréciée soit au niveau de l'ensemble des explicatives, soit individuellement.

- Tests de significativité de l'ensemble des coefficients : les trois statistiques usuelles, likelihood Ratio, Lagrange et Wald, sont affichées par PHREG. Ainsi, dans l'exemple qui clôt la section précédente, la table 4.11 est obtenue et on rejette naturellement la nullité simultanée des 8 coefficients.
- Tests de nullité individuels sur chaque coefficient : des tests de Wald sont présentés dans la table affichant les estimations [33]. PHREG sort également une table intitulée "Type 3 tests", qui pour les variables continues et les variables à deux modalités, dont évidemment une seule est présente dans la régression, ne fait que dupliquer les statistiques de la table précédente. Elle devient utile pour les variables entrées avec la commande `Class` ayant k modalités avec $k > 2$: dans ce cas, un test de Wald à $k-1$ degrés de liberté est présenté, l'hypothèse testée étant la nullité des coefficients afférents aux $k-1$ indicatrices présentes, *i.e.* que les $k-1$ modalités présentes ont le même impact que celle mise en référence, ou encore qu'il n'est pas utile de distinguer les individus selon les modalités de la variable en question.
- On pourrait également utiliser les résidus de Cox-Snell et refaire le test graphique préconisé après estimation d'un modèle paramétrique par la proc `LIFEREG`. Il est cependant généralement admis dans la littérature qu'ils ne sont pas d'une grande utilité avec le modèle de Cox. Comme nous allons le voir par la suite, dans ce modèle on va considérer d'autres résidus, qui pour certains sont des transformés de Cox-Snell.
- Pour comparer différents ajustements on dispose également des critères de sélection AIC d'Akaike et SBC de Schwarz.

4.6.2 Etude de spécification : résidus de martingales, régression locale et sommes partielles cumulées

Soit $N\_i(t)$ le nombre d'événements qu'a connu l'individu i sur la période $[0, t]$. Pour cet individu, le nombre d'événement prévus par le modèle est $H_i(t) = \exp(\hat{\beta^\top} x_i)\hat{H}_0(t)$. Pour des explicatives invariantes dans le temps, les résidus de martingale sont donnés par :

$$\hat{M}_i(t) = N_i(t) - H_i(t) \tag{4.42}$$
$$= N_i(t) - \exp(\hat{\beta} x_i)\hat{H}_0(t) \tag{4.43}$$
$$= N_i(t) + \log \hat{S}_i(t) \tag{4.44}$$

Ce résidu est donc simplement l'écart, mesuré sur une durée t et pour l'individu i, entre le nombre d'événements subis et le nombre d'événements prévus par l'équation ajustée.

33. Pour un exemple, voir la table 4.12.

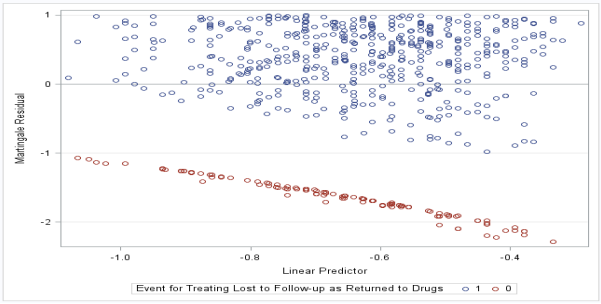


FIGURE 4.3 – Résidus de martingales versus index $\hat{\beta}^\top x$, individus censurés et non censurés

Dans un modèle de Cox $N_i(t)$ vaut typiquement 0 ou 1 et, compte-tenu du domaine de réalisation de $\hat{S}_i(t)$, les résidus de martingales appartiennent à $]-\infty, 1]$, leur distribution est donc fortement asymétrique. Positifs, ils sont le fait d'individus ayant connu l'événement trot tôt, négatifs ils signalent des individus qui survivent plus longtemps que prévu à l'événement [34]. Le graphe de ces résidus fait apparaître généralement deux nuages de points distincts qui distinguent individus censurés et non censurés. A des fins d'illustration, nous donnons dans la figure 4.3 le graphe de ces résidus obtenu en exécutant le code ci dessous, repris des exemples précédents qui demande la création de la table 'res' contenant en plus des données initiales les variables 'rmart' et 'index' contenant respectivement les résidus de martingales et $\hat{\beta}^\top x$.

```
proc phreg data=uis;
format treat prog.;
format site lieu.;
class treat(ref="court") site (ref='A') / param=ref;
model time*censor(0) = age treat site;
output out=res resmart=mart xbeta=index;
run;
quit;
proc sgplot data=res;
yaxis grid;
refline 0 / axis=y;
scatter y=mart x=index / group=censor;
run;
```

On peut montrer que ces résidus vérifient $E[M_i] = 0$ et $\text{cov}(M_i, M_j) = 0$, *i.e.* ils sont centrés et orthogonaux entre eux. Leur intérêt est d'aider au choix de la transformation à appliquer à une variable explicative continue [35] : soit x_1 cette variable, la question est de savoir si on doit l'intégrer, par exemple, comme x_1, $\log x_1$, $\sqrt{x_1}$, x_1^2,... Pour cela, deux démarches peuvent être mise en oeuvre :

1. **Régressions locales :** la première démarche exploite un résultat de Therneau, Grambsch et Flemming que l'on peut résumer comme suit : si l'index s'écrit sur une transformée de

34. Par exemple, si un individu i connaît l'événement à un temps t tel que $\hat{H}_i(t) = 5.4$, *i.e.* le modèle lui prévoit 5.4 événements à cette durée t_i, alors son résidu de martingale est $\hat{M}_i(t) = 1 - 5.4 = -4.4$.

35. Ils ne sont d'aucune utilité pour les variables catégorielles

x_1, et donc $h(t) = h_0(t)\exp[\beta_1 f(x_1)]$ où $f()$ est une fonction lisse, alors une régression non paramétrique, de type LOESS par exemple, des résidus de martingales tirés d'un ajustement où ne figure pas x_1, sur x_1 renseigne sur la fonction $f()$ dans la mesure où $E[M_i] = cf(x_{1i})$, c étant une constante dépendant de la proportion d'individus censurés. En pratique on réalise donc les étapes suivantes :

(a) ajustement d'une équation sans x_1, sauvegarde des résidus de martingales,

(b) régression LOESS de ces résidus sur x_1,

(c) graphe des résidus lissés en fonction de x_1 pour en tirer une indication visuelle sur la fonction $f()$,

(d) intégration de $f(x_1)$ dans la liste des explicatives. [36]

Afin d'illustrer cette démarche, nous reprenons le code précédent : on remarque que la variable 'ndrugtx' qui indique le nombre de traitements déjà subi par un individu avant d'intégrer l'échantillon de travail est absente de la liste des variables explicatives. Nous allons la considérer comme étant une variable continue. Les résidus de martingales obtenus en son absence ont été sauvegardés dans la table 'res' sous le nom 'mart'. Il suffit alors d'exécuter la régression locale :

```
proc loess data=res;
model mart=ndrugtx;
run;
```

La figure 4.4 est affichée. A sa vue, on peut raisonnablement retenir une fonction linéaire au moins après les 5 premières valeurs de 'ndrugtx'. En conséquence, on ajoute la variable en niveau à la liste des explicatives, on sauvegarde les nouveaux résidus de martingales et, si l'intégration par le niveau est satisfaisant, on doit maintenant obtenir une droite horizontale dans le graphiques des résidus lissés : la valeur de 'ndrugtx' ne devrait plus contenir d'information utile pour le résidu correspondant : $E[M\_i|\texttt{ndrugtx}_i] = E[M\_i] = 0$. Si on exécute :

```
proc phreg data=uis;
format treat prog.;
format site lieu.;
class treat(ref="court") site (ref='A') / param=ref;
model time*censor(0) = age treat site ndrugtx;
output out=res resmart=mart xbeta=index;
run;
quit;
proc loess data=res;
model mart=ndrugtx;
run;
```

On obtient alors la figure 4.5 dans laquelle le lissage s'est bien rapproché de l'horizontale. On remarque cependant une décroissance du lissage avec 'ndrugtx' qui laisse penser qu'une transformation concave, de type log(*ndrugtx*) est susceptible de faire mieux que le simple

36. Sachant que dans la plupart des études on rechigne à mettre des transformées non linéaires des variables comme explicatives en raison de la difficulté à interpréter la transformée en question. Par exemple, que signifie x^2, où $\sqrt{x}$? Pour cette raison on préfère souvent discrétiser la variable continue qui n'entrerait pas sous forme linéaire.

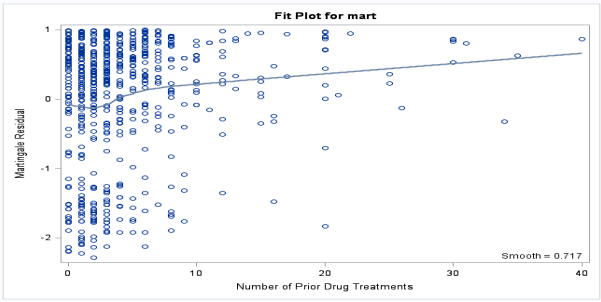


FIGURE 4.4 – Lissage des résidus de martingales par régression LOESS, variable NDRUGTX absente

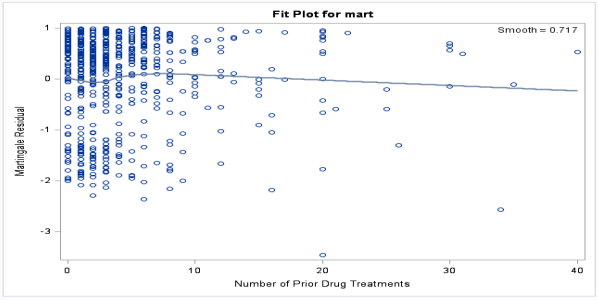


FIGURE 4.5 – Lissage des résidus de martingales par régression LOESS, variable NDRUGTX présente en niveau

niveau. L'exécution des commandes suivantes créée la nouvelle variable, l'intègre dans la liste des explicatives du modèle de Cox, sauvegarde les résidus de martingales, et estime la régression LOESS qui aboutit à la figure 4.6.

```
data uis2;
set uis;
lndrugtx=log(ndrugtx);
run;
proc phreg data=uis;
format treat prog.;
format site lieu.;
class treat(ref="court") site (ref='A') / param=ref;
model time*censor(0) = age treat site lndrugtx;
output out=res resmart=mart xbeta=index;
run;
quit;
proc loess data=res;
model mart=lndrugtx;
run;
```

La transformation logarithmique semble un peu plus appropriée que la prise du niveau de la variable 'ndrugtx'. Nous allons réexaminer ce choix de spécification en utilisant les sommes partielles de résidus de martingales.

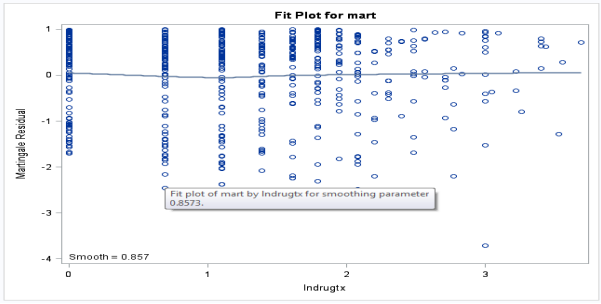


FIGURE 4.6 – Lissage des résidus de martingales par régression LOESS, variable LOG(NDRUGTX) présente

2. **Sommes partielles** de résidus de martingales : la commande ASSESS implémente les recommandations de Lin, Wei, and Ying. Pour une explicative continue quelconque, par exemple x_1, on peut construire la séquence de sommes partielles $\hat{W}_1(x)$ comme

$$\hat{W}_1(x) = \sum_{i=1}^{n} I(x_{1i} < x)\hat{M}_i \tag{4.45}$$

Si le modèle estimé est satisfaisant alors la séquence $\hat{W}_1(x)$ peut être approximée par un processus gaussien centré $\tilde{W}_1$ dont l'expression est complexe [37] mais qui peut être simulé. La commande ASSESS va grapher un certain nombre de ces processus simulés, 20 par défaut, et représenter sur le même graphe la suite observée $\hat{W}_1(x)$, en mettant en abscisse la variable x_1. A ce stade, l'aide à la spécification est seulement visuelle : l'explicative x_1 est considérée comme bien spécifiée si la trajectoire $\hat{W}_1(x)$ ne se démarque pas des trajectoires simulées $\tilde{W}_1(x)$, i.e. se situe dans la région du plan qui est balayé par les simulés. En revanche si $\hat{W}_1(x)$ est en dehors où trop proche des frontières supérieure où inférieure de cette région on sera amené à essayer une autre transformation sur l'explicative.
L'impression visuelle peut être confortée par le calcul d'un test de type Kolmogorov-Smirnov qui approxime au moyen de 1000 simulations de trajectoires $\tilde{W}_1$ la probabilité que $sup_x|\hat{W}_1(x)| \leq sup_x|\tilde{W}_1(x)|$. Naturellement, l'absence de rejet est favorable à la spécification retenue. Ce test est activé par le mot clef `RESAMPLE`.

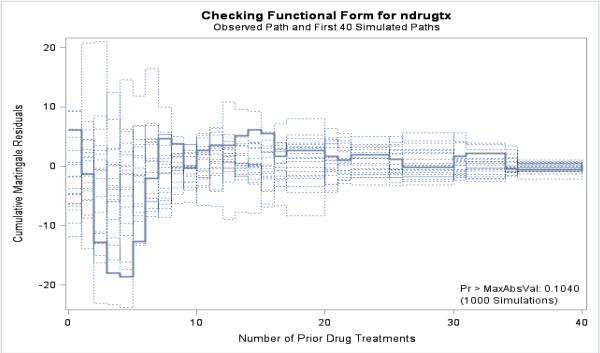
On illustre la démarche à suivre en reconsidérant le choix 'ndrugtx' versus 'log(ndrugtx)' abordé précédemment. Si nous insérons la commande ASSESS dans l'estimation faisant intervenir 'ndrugtx' en niveau

```
model time*censor(0) = age treat site ndrugtx;
assess var=(ndrugtx) npaths=40 resample;
```

On récupère le graphe et la table de la figure 4.7. Si on l'insère dans l'équation ayant log(ndrugtx) en explicative :

```
model time*censor(0) = age treat site lndrugtx;
assess var=(lndrugtx) npaths=40 resample;
```

37. Cf. l'aide de Proc PHREG, section Details - Assessment of the Proportional Hazards Model pour les curieux.



Supremum Test for Functional Form

Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
ndrugtx	18.5718	1000	604330632	0.1040

FIGURE 4.7 – Sommes partielles des résidus de martingales

Commande ASSESS - variable NDRUGTX présente en niveau

on obtient la figure 4.8.

Dans les deux cas, l'option `npath=40` réclame que 40 sommes simulées $\tilde{W}_1$ soient représentées. Dans cet example, il n'y a pas de réponse nettement privilégiée sur le plan visuel, et ce n'est qu'en prenant un seuil de risque de 10%-11% que le test KS permettrait de donner la préférence à la transformation en log.

4.6.3 Repérage des outliers : résidus de déviance, statistiques DFBETA et LD.

On sait que les résidus de martingales possèdent une distribution fortement asymétrique à gauche. En conséquence il est malaisé de répondre à une question telle que : un résidu de martingale fortement négatif est-il atypique où parfaitement raisonnable ? En raison de cette difficulté d'interprétation, pour détecter les outliers on va utiliser des transformés des résidus de martingales.

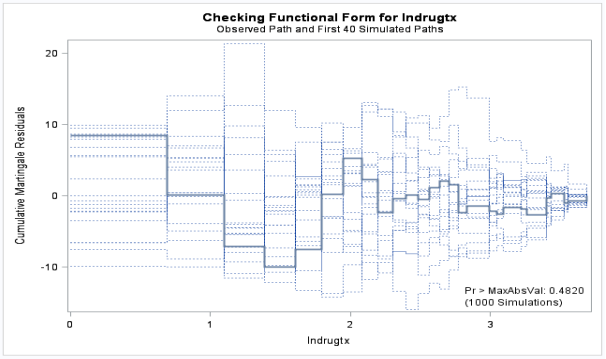
Les résidus de déviance :

Ils sont définisdéfinis comme :

$$\hat{D}_i(t) = sign[\hat{M}_i(t)] \sqrt{2[(-\hat{M}_i(t) - N_i(\infty)) \log \frac{N_i(\infty) - \hat{M}_i(t)}{N_i(\infty))}] \tag{4.46}$$

PHREG calcule un résidu de déviance par individu [38]. Dans cette transformée, les deux fonctions racine et log se conjuguent pour réduire la taille des résidus de martingale négatifs et augmenter celle des résidus positifs : on cherche ainsi à obtenir une distribution plus symétrique autour de zéro et possédant une variance unitaire. Leur interprétation est similaire à celles des résidus OLS dans un ajustement des moindres carrés : les individus ayant un résidu de déviance élevé en valeur absolue sont susceptibles d'être mal expliqués par l'équation ajustée et de correspondre à

38. En conséquence, ils ne sont pas calculés lorsque les données sont entrées selon un processus de comptage.



Supremum Test for Functional Form

Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
Indrugtx	9.9464	1000	1580808463	0.4820

FIGURE 4.8 – Sommes partielles des résidus de martingales

Commande ASSESS - variable Log(NDRUGTX) présente

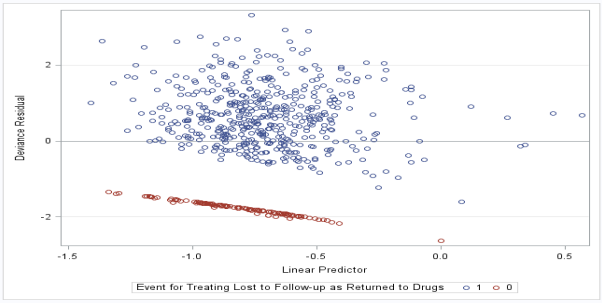


FIGURE 4.9 – Résidus de déviance versus index $\hat{\beta}^\top x$, individus censurés et non censurés

des outliers. Lorsque de tels individus sont détectés, la première chose à faire est de vérifier les données qui les concernent. Si ces dernières ne sont pas erronées, il faut décider de maintenir où non les individus en question dans l'échantillon de travail. Les statistiques DFBETA et LD peuvent alors aider à cette prise de décision.

Les résidus de deviance sont sauvegardés par l'emploi de l'option `resdev=` dans la commande `output`. Dans notre exemple pour les récupérer, avec les résidus de martingales et l'index estimé, dans un fichier dénommé 'res'. il suffit d'effectuer :

```
model time*censor(0) = age treat site ndrugtx;
output out=res resmart=mart xbeta=index resdev=dev;
```

Il est alors aisé d'obtenir la figure 4.9 et par exemple d'étudier plus spécifiquement les individus pour lesquels les résidus de déviance sont, en valeur absolue, supérieurs à 2. En comparant les figures 4.3 et 4.9 vous pouvez constater que ces individus atypiques sont effectivement plus aisément identifiables, à l'oeil nu, sur le résidus de déviance que sur les résidus de martingales.

Les statistiques DFBETA et LD :

Ces deux statistiques vont mesurer l'impact sur les résultats de l'estimation de chaque observation. La première observe les modifications des coefficients estimés des variables explicatives selon qu'une observation est inclue où non dans l'échantillon. la seconde évalue la variation de la valeur de la vraisemblance estimée.

1. Si on note $\hat{\beta}_{\setminus i}$ l'estimation de β obtenue lorsque la $i^{\text{ème}}$ observation est retirée de l'échantillon et $\hat{\beta}$ celle obtenue en sa présence, alors la statistique d'intérêt est égale à $DFBETA_i = (\hat{\beta}_{\setminus i} - \hat{\beta})$. C'est donc un vecteur a k composante pour chaque individu de l'échantillon.
   En regardant les composantes de DFBETA, il est évidemment possible d'avoir une information concernant l'influence de tout individu sur le coefficient estimé de chaque explicative, et donc de repérer ceux pour lequel cet impact apparaîtrait comme étant déraisonnable et donc susceptible d'être supprimé de la base de données. L'inconvénient majeur de cette démarche est qu'il faut réestimer le modèle autant de fois qu'il y a d'individus dans dans la base. Afin d'éviter cela, PHREG met en oeuvre une approximation qui évite les recalculs :

   $$DFBETA = \Im^{-1}(\hat{\beta})\hat{L}_i \tag{4.47}$$

   où $\hat{L}_i$ est l'estimation obtenue sur l'échantillon complet du résidu du score pour la $i^{\text{ème}}$ observation [39] .
   Un graphique par composante de DFBETA doit ensuite permettre de révéler les individus ayant le plus de poids dans l'estimation de chaque coefficient, d'avoir une appréciation de leurs impacts et en conséquence de permettre de statuer sur l'élimination ou non des individus concernés de la base de données utilisée pour l'estimation du modèle.
2. La statistique LD reprend la même logique que précédemment. Elle va simplement juger du poids d'un individu en évaluant la modification de la log-vraisemblance qu'entraîne son retrait de l'échantillon. On peut la percevoir comme une statistique donnant une information plus globale que DFBETA qui s'intéressait au coefficient de chacune des explicatives.
   Avec les notations précédentes, LD mesure la variation de la vraisemblance par :

   $$LD_i = 2[L(\hat{\beta}_{\setminus i}) - L(\hat{\beta})] \tag{4.48}$$

   L'inconvénient est, comme précédemment, que son calcul requiert un nombre d'estimations du modèle égal à la taille de l'échantillon. Pour éviter ces calculs, PHREG fait encore appel à une approximation :

   $$LD_i = \hat{L}_i{}^{\top}\Im^{-1}(\hat{\beta})\hat{L}_i \tag{4.49}$$

Ces statistiques sont sauvegardées dans la table référencée dans la commande `output=` au moyen des options

— `DFBETA` = *liste de noms* en nombre inférieur ou égal au nombre d'explicatives de la commande `model`, k. Si la taille de la liste, k_0 est inférieure à k, seules les variations des k_0 premiers coefficients sont sauvegardées.

39. Important : lorsque les données sont entrées selon les modalités des processus de comptage, *i.e.* plus d'un enregistrement par individu, alors les résidus du score et, par voie de conséquence, DFBETA, sont évalués par enregistrement. En conséquence, pour retrouver la mesure DFBETA propre à un individu donné, il faut additionner les évaluations partielles le concernant en passant par une étape PROC MEANS comme cela est fait dans l'exemple '64.7 Time-Dependent Repeated Measurements of a Covariate' dans la documentation de PHREG.

— `LD=` *nom de variable*. Cette variable contiendra les variations de la fonction de vraisemblance associé au retrait de chaque observation.

4.6.4 Tests de l'hypothèse PH - Introduction d'interactions avec le temps, Résidus de Schoenfeld et sommes de transformées de résidus de martingale

Test de l'hypothèse PH par introduction d'interactions avec le temps :

La facilité de construction par programme de variables dépendantes des durées permet de construire aisément un test de l'hypothèse PH sur les explicatives invariantes dans le temps où dont les écarts de valeurs entre individus différents ne changent pas avec les durées. Pour en comprendre la logique, considérons un cadre très simple, avec une seule explicative x_1. L'hypothèse PH correspond au fait que si x_1 est invariant dans le temps [40], alors :

$$\frac{h_j(t)}{h_i(t)} = \frac{h_0(t)\exp(\beta_1 x_{1j})}{h_0(t)\exp(\beta_1 x_{1i})} = \exp(\beta_1[x_{1j} - x_{1i}]) = \text{constante} \tag{4.50}$$

L'objectif du test sera de vérifier cette constance lorsque $x_{1j} \neq x_{1i}$. Pour cela on ajoute dans les explicatives une variable d'interaction entre le temps et x_1, par exemple le produit $x_1 \times t$. Il vient alors :

$$\frac{h_j(t)}{h_i(t)} = \frac{h_0(t)\exp(\beta_1 x_{1j} + \beta_2 x_{1j}\, t)}{h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{1i}\, t)} = \exp(\beta_1[x_{1j} - x_{1i}] + \beta_2[x_{1j} - x_{1i}]t) \tag{4.51}$$

Il est alors évident que dans cette version augmentée, on ne retrouve le résultat 4.50, c'est à dire la constance du ratio de risque uniquement si $\beta_2 = 0$. Lorsque ce n'est pas le cas, le ratio se déforme avec les durées ce qui invalide alors l'hypothèse PH pour l'explicative x_1. En pratique, dans ce dernier cas on devrait laisser la variable d'interaction, $x_1 \times t$, dans la liste des explicatives

On peut illustrer la démarche en reprenant le fichier `'uis'` utilisé dans les précédents exemples. On va maintenant tester la validité de l'hypothèse de risque proportionnel pour les variables 'treat', qui distingue entre traitement court et traitement long, 'age', et 'ndrugtx' qui indiquent respectivement l'âge de l'individu lors de l'entrée dans le programme de soin étudié et le nombre de traitements dont a bénéficié un individu avant son incorporation dans ce programme. Les commandes d'appel de `PHREG` pourraient alors ressembler à [41] :

```
proc phreg data=uis;
format treat prog.;
format site lieu.;
class treat(ref="court") site(ref='A') / param=ref;
model time*censor(0) = age ndrugtx treat site age_logt ndrugtx_logt treat_logt
site_logt;
age_logt = age*log(time);
ndrugtx_logt = ndrugtx*log(time);
treat_logt = treat*log(time);
```

40. En fait, l'important est que les écarts $x_{1j} - x_{1i}$ soient constants dans le temps. Cela est évidemment vrai si x_1 est une constante pour chaque individu mais ce n'est pas une condition nécessaire pour la réalisation du test.

41. SAS conseille, pour des raisons numériques, d'utiliser une interaction non pas avec t, mais avec $\log t$, ici nous avons essayé les deux versions.

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age		1	-0.02539	0.03400	0.5575	0.4553	0.975	Age at Enrollment
ndrugtx		1	0.01892	0.03199	0.3497	0.5543	1.019	Number of Prior Drug Treatments
treat	long	1	-0.66067	0.41139	2.5790	0.1083	0.517	Treatment Randomization Assignment long
site	B	1	-0.58085	0.47053	1.5239	0.2170	0.559	Treatment Site B
age_logt		1	0.0006073	0.00712	0.0073	0.9320	1.001	
ndrugtx_logt		1	0.00361	0.00692	0.2719	0.6020	1.004	
treat_logt		1	0.08982	0.08629	1.0834	0.2979	1.094	
site_logt		1	0.08865	0.09797	0.8189	0.3655	1.093	

TABLE 4.12 – Test de l'hypothèse PH - Variables d'interaction avec les durées

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age		1	-0.02432	0.01225	3.9395	0.0472	0.976	Age at Enrollment
ndrugtx		1	0.02356	0.01229	3.6729	0.0553	1.024	Number of Prior Drug Treatments
treat	long	1	-0.52604	0.14922	12.4279	0.0004	0.591	Treatment Randomization Assignment long
site	B	1	-0.32346	0.16721	3.7421	0.0531	0.724	Treatment Site B
age_t		1	5.58904E-6	0.0000607	0.0085	0.9266	1.000	
ndrugtx_t		1	0.0000801	0.0000628	1.6299	0.2017	1.000	
treat_t		1	0.00181	0.0007603	5.6451	0.0175	1.002	
site_t		1	0.00100	0.0008338	1.4460	0.2292	1.001	

TABLE 4.13 – Test de l'hypothèse PH - Variables d'interaction avec les durées

```
site_logt=site*log(time);
run;
quit;
```

Les résultats de la table 4.12 sont alors obtenus. Lorsque le temps n'est pas transformé par passage aux logarithmes, nous arrivons aux résultats de la table 4.13. Aux seuils usuels de risque, nous ne pouvons rejeter la validité de l'hypothèse PH pour les trois variables 'age', 'ndrugtx' et 'site'. La conclusion diffère pour la durée du traitement selon qu'une transformation logarithmique est appliquée (dans ce cas, non rejet de PH) ou non (rejet de PH). Si on veut prendre ce dernier résultat en compte, il faudrait alors laisser la variable `treat*time` parmi les explicatives.

Les résidus de Schoenfeld :

Soit t_i le temps de survenu de l'événement étudié pour l'individu i, soit x_{ij} la valeur de la $j^{\text{ième}}$ explicative pour cet individu, et soit $\Re_{t_i}$ l'ensemble des individus à risque en t_i. Le résidu de Schoenfeld associé à l'individu i et à la $j^{\text{ième}}$ explicative, est donné par :

$$\hat{s}_{ij} = x_{ij} - \sum_{r \in \Re_{t_i}} x_{rj} \hat{p}_{r,t_i}, \text{ avec} \tag{4.52}$$

$$\hat{p}_{r,t_i} = \frac{\exp \hat{\beta}^\top \mathbf{x}_r}{\sum_{r \in \Re_{t_i}} \exp \hat{\beta}^\top \mathbf{x}_r} \tag{4.53}$$

où $\hat{p}_{r,t_i}$ est la vraisemblance estimée que l'individu r à risque connaisse l'événement en t_i. Le résidu de Schoenfeld est donc un écart entre la valeur d'une explicative observée sur un individu qui connaît l'événement à une certaine durée et une moyenne pondérée des valeurs de cette explicative observées à cette même durée sur tous les individus alors à risque.
On notera que ces résidus ne sont définis que pour les individus non censurés, et pour chacune

des explicatives : si on a k explicatives, on associe à chaque individu un vecteur de k résidus de Schoenfeld, *i.e.*,

$$\hat{\mathbf{s}}_i = (\hat{s}_{i1}, \hat{s}_{i2}, \ldots, \hat{s}_{ik})^\top$$

Leur intérêt provient du fait qu'ils sont fonction de l'écart entre les coefficients β_j, $j = 1, \ldots, k$ du modèle de Cox et les coefficients $\beta_{j,t}$, $j = 1, \ldots, k$ d'un modèle de Cox à coefficients variables avec les durées $E[\hat{s}_{ij}] = \beta_{j,t_i} - \beta_j$ En conséquence, pour chaque explicative xj, $j = 1 \ldots, k$, ils vont permettre de statuer, au moins visuellement, sur le test $H0 : \beta_{j,t_i} = \beta_j$ versus $H1 : \beta_{j,t_i} \neq \beta_j$, *i.e.* constance versus non constance des coefficients en fonction des durées et donc sur la validité de l'hypothèse PH, celle-ci supposant leur constance.
Pour cela, on va grapher les résidus de Schoenfeld en fonction des durées, avec donc un graphe par explicative. Sous $H0$, ils devraient être aléatoirement distribués autour de zéro. Pour faciliter la lecture des graphes, on préfère travailler avec les résidus de Schoenfeld standardisés :

$$\mathbf{r}_i = n_e \Im^{-1}(\hat{\beta}) \hat{\mathbf{s}}_i \tag{4.54}$$

où $\Im^{-1}(\hat{\beta})$ est l'estimation de la matrice de variance-covariance des coefficients $\hat{\beta}$ et n_e le nombre total d'événements observés dans l'échantillon.
La sauvegarde des résidus de Schoenfeld s'effectue au sein de la commande `OUTPUT` au moyen de l'option `RESSCH=` $\texttt{nom}_1$ $\texttt{nom}_2 \ldots$ $\texttt{nom}_k$, et pour les résidus de Schoenfeld normalisés, par `WTRESSCH=` $\texttt{nom}_1$ $\texttt{nom}_2 \ldots$ $\texttt{nom}_k$. Notez qu'avec une série de résidus par explicative, il faut préciser une liste de noms pour ces résidus. On retrouve ici les mêmes règles qu'avec les statistiques DFBETA : s'il y a k explicatives dans l'équation estimée et si la longueur de la liste de noms est $k_1 < k$, alors seuls les résidus associés aux k_1 première explicatives de la commande `MODEL` sont sauvegardés.
On peut ensuite ajuster une régression locale sur chaque série de résidus, l'hypothèse nulle étant favorisée par l'obtention d'une horizontale. Par ailleurs, la forme du lissage peut servir d'indication quant à la façon dont le coefficient varie avec les durées.
On illustre la démarche avec les données de l'exemple précédent et en nous intéressant à la constance des coefficients de `age` et `ndrugtx`. L'appel à PHREG pourrait ainsi incorporer les lignes suivantes [42] :

```
model time*censor(0) = age ndrugtx treat site;
output out=res wtressch=wtsch_age wtsch_lndrugtx;
```

et se poursuivre par :

```
proc loess data=res;
model wtsch_age=time;
run;
proc loess data=res;
model wtsch_ndrugtx=time;
run;
```

On obtient finalement les graphes 4.10 et 4.11

Dans cet exemple, on ne constate pas de déviation importante par rapport à l'horizontale, ce qui joue en faveur du respect de l'hypothèse PH pour les deux variables étudiées.

42. Pensez alors à mettre les explicatives `age` et `ndrugtx` en première et seconde position dans la commande `model`.

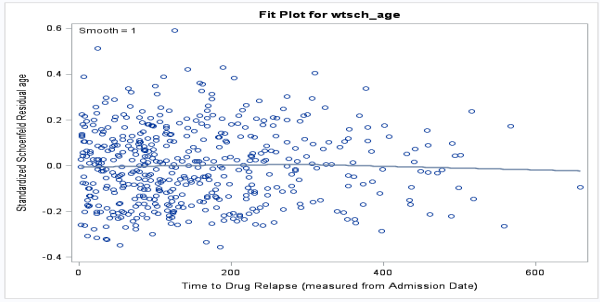


FIGURE 4.10 – Résidus de Schoenfeld normalisés associés à `age`

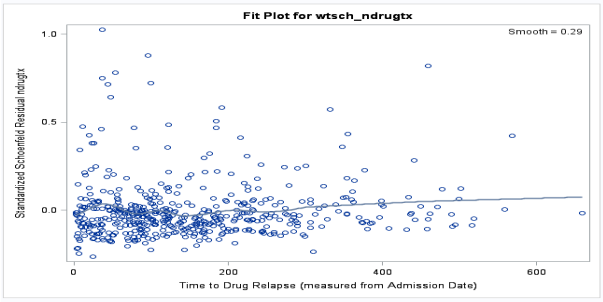


FIGURE 4.11 – Résidus de Schoenfeld normalisés associés à `ndrugtx`

Les sommes de transformées de résidus de martingales :

A chaque date d'événement t on peut définir les valeurs d'un processus $\underset{k,1}{\hat{\mathbf{U}}_t}$ comme

$$\hat{\mathbf{U}}(t) = \sum_{i=1}^{n} \mathbf{x}_i \hat{M}_i(t) \tag{4.55}$$

où $\mathbf{x}_i = (x_{i1}, x_{i1}, \ldots, x_{ik})^\top$ contient les valeurs des caractéristiques de l'individu i mesurées sur les k explicatives du modèle, et $\hat{M}_i(t)$ la valeur estimée du résidu de martingale pour cet individu au temps t dont les éléments standardisés sont donnés par :

$$\hat{U}_{c_j} = \left[\Im^{-1}(\hat{\beta})_{jj}\right]^{-1/2} \hat{U}_j(t), j = 1, \ldots, k \tag{4.56}$$

Sous l'hypothèse nulle de validité de l'hypothèse PH, ces processus standardisées peuvent être approximés par des processus gaussiens centrés $\tilde{U}_{c_j}, j = 1, \ldots, k$. L'idée est de comparer la trajectoire observée sur les $\hat{U}_{c_j}$ à un certain nombre de trajectoires simulées $\tilde{U}_{c_j}$. Sous H0=validité de l'hypothèse PH, la trajectoire observée doit se fondre dans les trajectoires simulées. Comme dans lors de l'étude de spécification discuté précédemment, l'information visuelle est confortée par un test de type KS tel qu'un significance level inférieur au seuil de risque choisit doit conduire au rejet de l'hypothèse PH.
Ces graphes et tests sont réalisés pour chacune des variables explicatives via la commande `ASSESS` dans laquelle il faut simplement faire apparaître le mot clef `PH` pour que ce soit cette hypothèse PH qui fasse l'objet de la commande.

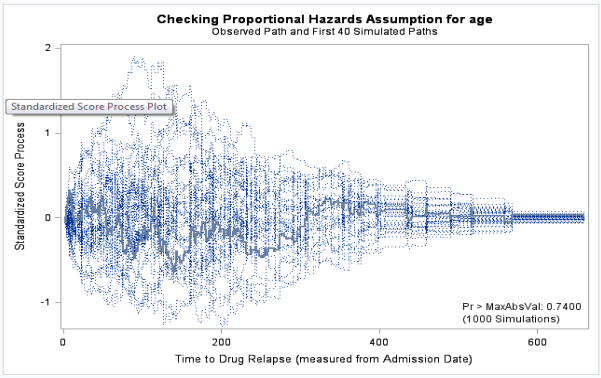


FIGURE 4.12 – Commande ASSESS, Test hypothèse PH - variable age

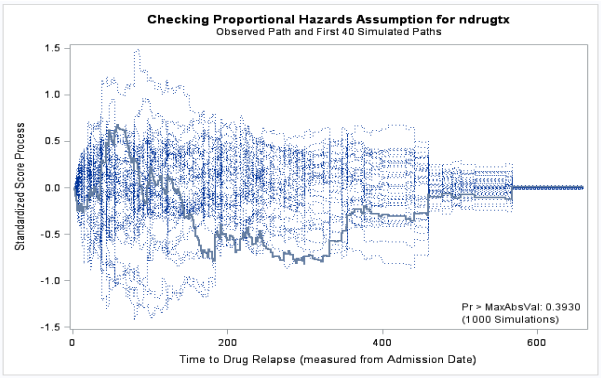


FIGURE 4.13 – Commande ASSESS, Test hypothèse PH - variable NDRUGTX

Pour illustrer la démarche, nous allons reconsidérer l'hypothèse PH sur toutes les explicatives du dernier programme. Il suffit donc de lui ajouter la ligne

```
assess ph / npaths=40 resample;
```

On récupère les graphes 4.12, 4.13, 4.14 et 4.15, dans chacun desquels 40 trajectoires simulées sont représentées en plus de l'observée, ainsi que le test KS généré par l'option `resample`. En ce qui concerne ce dernier, la table récapitulative 4.14 est également fournie. On constate que cette méthode ne remet pas en cause l'hypothèse de risques proportionnels sauf pour la variable relative à la durée du traitement, confirmant ainsi un résultat antérieur (Cf. table 4.13) obtenu à l'aide de la prise en compte d'interactions avec les durées.

4.7 La sélection automatique des variables explicatives

Avec PHREG il est possible de sélectionner automatiquement les variables explicatives à retenir dans l'estimation finale du modèle parmi une liste de variables candidates. Si cette particularité semble intéressante pour construire rapidement le modèle final, il faut toujours avoir à l'esprit ses insuffisances. La plus importante est l'absence complète de référence à la théorie dans le choix du modèle retenu. L'interprétation des résultats obtenus peut donc s'avérer particulièrement délicate

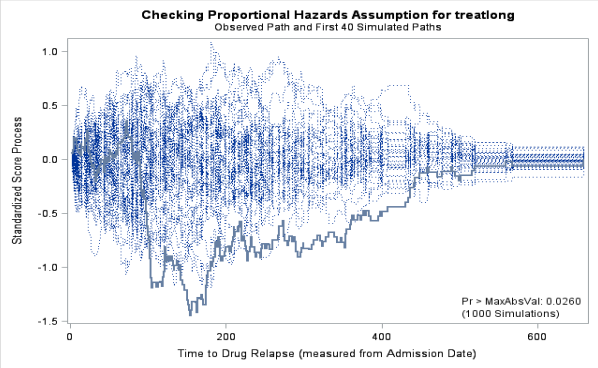


FIGURE 4.14 – Commande ASSESS, Test hypothèse PH - modalité 'long' de la variable Treat

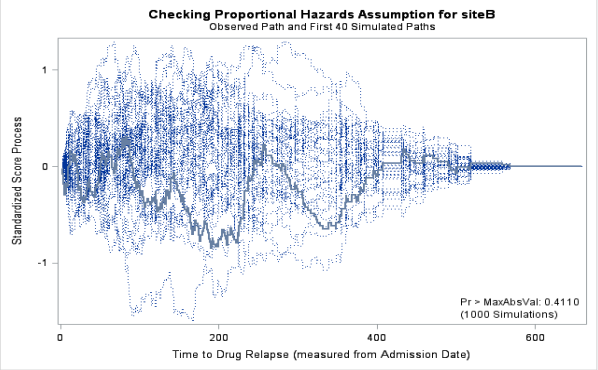


FIGURE 4.15 – Commande ASSESS, Test hypothèse PH - modalité B de la variable Site

Supremum Test for Proportionals Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
age	0.6367	1000	501972495	0.7400
ndrugtx	0.8253	1000	501972495	0.3930
treatlong	1.4456	1000	501972495	0.0260
siteB	0.8464	1000	501972495	0.4110

TABLE 4.14 – Commande ASSESS - récapitulatif des tests KS

puisque seules les propriétés statistiques des observations sont utilisées. Par ailleurs rien n'assure la robustesse du modèle sélectionné : il suffit de changer d'échantillon pour qu'éventuellement la liste des explicatives retenues soit complètement bouleversée. Au minimum il est donc utile que les variables candidates aient elles-mêmes fait l'objet d'une sélection fondée théoriquement.

Pour autant ces méthodes ne manquent pas complètement d'intérêt. Supposons que plusieurs variables soient en théorie a priori pertinentes pour expliquer la fonction de risque mais qu'existent entre elles de fortes corrélations de sorte que l'on ne puisse pas déterminer théoriquement quel sous-ensemble suffit à l'explication du phénomène et quelles variables, conditionnellement à la présence d'explicatives déjà retenues, peuvent être délaissées. Dans ce cas de figure il ne semble pas déraisonnable de fonder le choix sur des techniques de sélection automatique. Il est cependant possible que différentes techniques conduisent à des sélections différentes. PHREG en propose cinq dont seulement quatre nous intéressent ici puisque l'option `SELECTION=NONE`, qui est prise par défaut, réclame l'estimation du seul modèle contenant la totalité des explicatives spécifiées par l'utilisateur. Les quatre autres sont FORWARD, BACKWARD, STEPWISE, et SCORE.

1. SELECTION=FORWARD : PHREG estime le modèle avec constante et les premières k_0 explicatives de la liste de k variables, où k_0 est fixé par l'option START=k_0 ou INCLUDE=k_0. Par défaut $k_0 = 0$. Ensuite la procédure recherche parmi les variables restantes la plus significative et l'ajoute au modèle si son seuil de significativité est inférieur au seuil fixé par SLENTRY= . Une fois entrée dans le modèle la variable n'est jamais retirée. La démarche est reprise avec le modèle à $k_0 + 1$ explicatives. Elle s'arrête lorsque la plus significative des variables non encore incorporées a un seuil de significativité supérieur à la valeur exigée par SLENTRY.
2. SELECTION=BACKWARD : la procédure estime le modèle ayant la totalité des explicatives (ou seulement les k_0 premières si l'option START=k_0 est utilisée). A l'aide d'un test de Wald, la moins significative est retirée dès lors que son seuil de significativité est supérieur à la valeur exigée par SLSTAY=. La procédure arrête lorsque plus aucune variable n'est autorisée à sortir.
3. SELECTION=STEPWISE : la procédure s'exécute comme avec l'option FORWARD à la différence près q'une variable entrée à une étape de la sélection peut sortir du modèle si, à une étape ultérieure et donc après prise en compte de nouvelles explicatives, son seuil de significativité passe au-dessus de la valeur requise par SLSTAY.
4. SELECTION=SCORE. PHREG recherche pour un nombre d'explicatives fixé (par défaut 1, puis 2, etc.) les "meilleurs" modèles au sens du test de Lagrange de nullité des coefficients. Les nombres d'explicatives minimal et maximal sont gérés respectivement par les options START= et STOP= . Le nombre de "meilleurs" modèles est géré par l'option BEST= . Par exemple, BEST=2 START=3, STOP=5 recherche les 2 meilleurs modèles dans la totalité des modèles possibles à 3, puis 4, puis 5 explicatives prises parmi la liste initiale à k éléments. L'option STOP gérant la taille maximale du modèle peut être utilisée dans tous les modes de sélection.

Chapitre 5

l'ajustement du risque en temps discret

En temps discret on va trouver essentiellement deux modélisations. L'une est une régression logistique qui peut être vue comme une approximation sous certaines conditions du modèle de Cox en temps continu, l'autre basée sur une fonction de lien log-log en est le pendant exact en ce sens qu'elle respecte strictement l'hypothèse de risque proportionnel.
Ces deux formulations peuvent donc être vues comme des adaptations du modèle de Cox en temps continu à des temps discrets. Elles ont cependant leur intérêt propre. C'est évidemment le cas lorsque le temps qui s'applique au phénomène étudié est effectivement discret. C'est également le cas lorsque le nombre de périodes de suivi des individus est faible : il est préférable dans ce cas de supposer un temps discret. C'est encore le cas lorsque l'on veut préciser la fonction de risque et son évolution avec les durées : dans le modèle en temps continu, cette fonction est relativement secondaire car ce qui importe est surtout l'estimation de l'impact des caractéristiques des individus sur le risque plutôt que le risque lui-même. Enfin c'est aussi le cas lorsque le nombre d'événements concomitants est élevé. On connaît les difficultés rencontrées alors par le modèle en temps continu, celles-ci disparaissant lorsqu'on est en temps discret comme nous allons le voir.

5.1 L'écriture du modèle

5.1.1 la régression logistique

On sait qu'en temps discret le risque est une probabilité et est donc compris entre 0 et 1. Cox a proposé dans ce cas de recourir à un ajustement de type logistique. Si le vecteur x_i contient les caractéristiques d'un individu i , alors son risque à une durée t serait donc donné par :

$$h_i(t) = \frac{exp(\beta_{0t} + x^\top \beta)}{1 + exp(\beta_{0t} + x^\top \beta)} = \frac{1}{1 + e^{-(\beta_{0t} + x^\top \beta)}} \tag{5.1}$$

conduisant à l'équation logistique :

$$\log(\frac{h(t)}{1 - h(t)}) = \text{logit}(h(t)) = \beta_{0t} + x^\top \beta \tag{5.2}$$

soit encore :

$$\frac{h(t)}{1 - h(t)} = \exp(\beta_{0t}) \times \exp(x^\top \beta) = \frac{h_0(t)}{1 - h_0(t)} \exp(x^\top \beta) \tag{5.3}$$

où $\beta_{0t} = \text{logit}(h_{0t})$, et h_{0t} serait le risque de base ne dépendant que de la durée et pas des caractéristiques des individus.

On notera toutefois que dans cette formulation l'hypothèse centrale du modèle de Cox, à savoir celle de risques proportionnels, ne tient plus [1]. Si on veut vérifier cette hypothèse, il faut prendre une fonction de lien non pas logistique mais de type $log - log$ entre le risque et l'index des caractéristiques individuelles.

5.1.2 la fonction de lien log-log ou complementary log-log

L'égalité suivante est toujours valide :

$$S_i(t) = (1 - h_i(t)) * S_i(t-1)$$

Elle rappelle simplement que pour connaître un événement après une durée t, il faut qu'il ne se soit pas réalisé avant t et, conditionnellement à cela, qu'il ne se soit pas non plus réalisé en t. Par substitutions répétées et en se rappelant que $S(0) = 1$, on arrive à :

$$S_i(t) = [1 - h_i(t)][1 - h_i(t-1)] \ldots [1 - h_i(1)]$$

On sait par ailleurs, voir équation (3.1.2) page 64, que lorsque l'hypothèse de risques proportionnels est valide, alors la fonction de survie d'un individu possédant les caractéristiques x_i est reliée à la survie de base selon :

$$S_i(t) = S_0(t)^{exp(x_i^\top \beta)}$$

En combinant les deux dernières égalités, on montre aisément que sous hypothèse de validité de l'hypothèse PH, il vient :

$$[1 - h_i(t)] = [1 - h_0(t)]^{exp(x_i^\top \beta)}$$

et finalement :

$$\log[-\log[1 - h_i(t)]] = x_i^\top \beta + \beta_{0t} \tag{5.4}$$

ou si on préfère :

$$h_i(t) = 1 - \exp\left[-exp[x_i^\top \beta + \beta_{0t}]\right] \tag{5.5}$$

avec $\beta_{0t} = \log(-\log[1 - h_0(t)])$. Ainsi la transformée $\log - \log$ du risque nous donne une équation qui respecte l'hypothèse PH. On peut par ailleurs montrer qu'ici les coefficients estimés seront identiques quelle que soit la longueur de l'intervalle de temps utilisée (mois, trimestre,...) alors que leurs valeurs vont se modifier en fonction du choix fait sur cette longueur d'intervalle dans le cas de la régression logistique. Allison remarque cependant qu'en pratique les deux régressions

1. Il suffit de faire le rapport des risques de deux individus i et j pour s'en apercevoir. Vous pouvez le vérifier à titre d'exercice. Cependant cette formulation logistique approximerait bien le modèle de Cox en temps continu lorsque la largeur des intervalles sur lesquels le temps est mesuré tend vers zéro.

donnent des résultats souvent proches [2, 3].

Pour clore ce point, remarquez qu'avec les équations (5.2) et (5.4), ces deux versions appartiennent à la classe des modèles linéaires généralisés [4].

5.2 Son estimation

Pour son estimation nous allons considérer sa fonction de vraisemblance. Pour mémoire on ne considère ici que la possibilité de censure à droite. Dans ces conditions, la durée reportée pour un individu i quelconque, t_i, correspond soit à la réalisation de l'événement étudié soit à une censure à droite. Tout individu va donc entrer dans la vraisemblance soit avec la probabilité que l'événement se réalise en t_i, soit avec la probabilité qu'il se réalise après t_i. Si on considère l'indicatrice $\delta_i = 0$ en cas de censure et $\delta_i = 1$ sinon, alors la vraisemblance s'écrit :

$$L = \prod_{i=1}^{n} [Prob(T_i = t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}$$

Comme

$$\begin{aligned} Prob(T_i = t_i) &= Prob(T_i = t_i \wedge T_i > t_i - 1) \\ &= Prob(T_i = t_i \,|\, T_i > t_i - 1) \times Prob(T_i > t_i - 1) \\ &= h(t_i) \times [1 - h_i(1)][1 - h_i(2)] \dots [1 - h_i(t_i - 1)], \text{ et,} \qquad (5.6) \\ S(t_i) &= [1 - h_i(1)][1 - h_i(2)] \dots [1 - h_i(t_i)] \qquad (5.7) \end{aligned}$$

Il vient :

$$\begin{aligned} L &= \prod_{i=1}^{n} \{h(t_i) \times [1 - h_i(1)][1 - h_i(2)] \dots [1 - h_i(t_i - 1)]\}^{\delta_i} \{[(1 - h_i(1)][(1 - h_i(2)] \dots [1 - h_i(t_i)]\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} h(t_i)^{\delta_i} [1 - h_i(t_i)]^{1-\delta_i} [(1 - h_i(1)][(1 - h_i(2)] \dots [1 - h_i(t_i - 1)] \end{aligned}$$

et donc une log-vraisemblance égale à :

$$\ell = \sum_{i=1}^{n} \left(\delta_i \log \frac{h(t_i)}{1 - h_i(t_i)} + \sum_{j=1}^{t_i} \log[(1 - h_i(j)] \right)$$

2. Dans un travail récent, Hee-Koung Joeng, Ming-Hui Chen, et Sangwook Kang concluent que cette proximité des résultats est d'autant plus grande que le risque de survenue de l'événement étudié est faible et notamment inférieur à 0.08. Dans ce travail, ils comparent également les critères de sélection AIC et BIC pour ce qui concerne le choix de la fonction de lien et le choix du modèle optimal et conseillent l'emploi du critère d'Akaike, "Proportional exponentiated link transformed hazards (ELTH) models for discrete time survival data with application", *Lifetime Data Anal.*, January ; 22(1) : 38-62, 2016.

3. On retiendra plus souvent le modèle Logit suivant en cela l'opinion de Stephen Jenkins : "there is no particular reason why, with economic duration data, hazards should be proportional."(p. 134), in S. P. Jenkins, Easy estimation for discrete-time duration models, *Oxford Bulletin of Economic and Statistics*, 57, 1 , 129-138, 1995.

4. Pour mémoire, le modèle linéaire d'école, $y = X\beta + \epsilon$ est tel que $E[y] = \mu_y = X\beta$, *i.e.* l'espérance de l'expliquée est linéaire en β. Dans un modèle linéaire généralisé, c'est la transformée de l'espérance de l'expliquée par une fonction différentiable $f()$, dite fonction de lien, qui aura une écriture linéaire sur β, soit $f(\mu_y) = X\beta$.

On peut à ce stade envisager de maximiser cette log-vraisemblance après avoir donné à $h_i(t_i)$ l'expression donnée par l'équation (5.1) si on a fait le choix de la régression logistique, ou celle donnée par (5.5) si on préfère utiliser la transformée log-log.

On doit cependant à Brown [5] et Allison [6] une ré-écriture de cette vraisemblance qui va grandement simplifier l'estimation de ce modèle en effaçant la complexité qu'introduit la distinction censure-non censure de même que celle liée aux événement concomitants et justifiant l'emploi d'algorithmes d'estimation connus. L'idée de Brown reprise et popularisée par Allison est de créer pour chaque individu une indicatrice y_{it} telle que $y_{it} = 1$ si l'événement s'est réalisé à la durée t pour l'individu i et $y_{it} = 0$ sinon. Prenons une personne qui a connu l'événement au bout de 3 trimestres, les observations de cette indicatrice seront simplement 0,0,1.
Si on écrit la log-vraisemblance non pas sur l'indicatrice de censure δ_i comme précédemment, mais sur cette nouvelle indicatrice y_{it}, il vient :

$$\ell = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \left(y_{it} \log \frac{h(t_i)}{1 - h_i(t_i)} + \log[(1 - h_i(j)] \right) \tag{5.8}$$

et sous cette forme vous devez reconnaître l'écriture de la log-vraisemblance des modèles à variable expliquée y dichotomique.
Grâce à cette équivalence on peut donc utiliser les procédures d'estimation de variables expliquées qualitatives pour ajuster le modèle de durée à temps discret et notamment la procédure bien connue de régression logistique si l'on opte pour la fonction de lien qui lui correspond, *i.e.* si le risque est modélisé selon l'équation (5.1) qui de ce fait est devenue la méthode la plus usitée dans ce cadre de temps discret. Outre la facilité de mise en oeuvre, elle offre également d'autres avantages comme une prise en compte aisée de variables explicatives dépendantes du temps, mais aussi d'interactions avec le temps, *i.e.* d'avoir des explicatives dont l'impact se modifie avec la durée. L'inconvénient de cette procédure est en général l'obligation de transformer la table des observations initiales, qui présente souvent un seul enregistrement par individu, en une table où à chaque individu va correspondre un nombre d'enregistrements égal au nombre de périodes sur lesquelles il est observé.

Avant de donner un exemple de transformation de table, on va préciser le traitement afférent au risque de base. On sait que ce risque est implicite au coefficient β_{0t} de nos équations définissant le risque instantané h_{it}. L'approche la plus commune est de supposer que ce risque de base, et donc ce coefficient, se modifie dans le temps. On peut par exemple opter pour une formulation telle que $\beta_{ot} = \alpha_0 + \alpha_1 t$, ou tout autre polynôme sur le temps. Toutefois, en l'absence d'a priori, on retient généralement de le représenter de façon non contraignante par une constante qui peut varier de période en période, obligeant alors simplement à la création d'indicatrice de chacune des périodes.

Afin d'illustrer ce qui précède, nous allons considérer un fichier de données initial qui contiendrait les trois individus dont les caractéristiques sont décrites par deux variables x_1 et x_2 comme présentées dans la table 5.1. Pour simplifier, on a supposé que ces 3 individus étaient entrés à la

5. C. C. Brown, On the Use of Indicator Variables for Studying the Time Dependance of Parameters in a Response-Time Model, *Biometrics* 31, pp. 863-872, 1975.

6. Paul D. Allison, Discrete-Time Methods for the Analysis of Event Histories, *Sociological Methodology*, vol. 13, pp. 61-98, 1982.

Id	durée	cens	x_1	x_2
1	3	1	x_{11}	x_{12}
2	2	0	x_{21}	x_{22}
3	4	1	x_{31}	x_{32}

TABLE 5.1 – Exemple de fichier de données initial

Id	y	x_1	x_2	T_1	T_2	T_3	T_4
1	0	x_{11}	x_{12}	1	0	0	0
1	0	x_{11}	x_{12}	0	1	0	0
1	1	x_{11}	x_{12}	0	0	1	0
2	0	x_{21}	x_{22}	1	0	0	0
2	0	x_{21}	x_{22}	0	1	0	0
3	0	x_{31}	x_{32}	1	0	0	0
3	0	x_{31}	x_{32}	0	1	0	0
3	0	x_{31}	x_{32}	0	0	1	0
3	1	x_{31}	x_{32}	0	0	0	1

TABLE 5.2 – Fichier d'exemple retraité pour l'estimation en temps discret avec un risque de base de montant spécifique à chaque intervalle de temps.

même date dans l'échantillon de travail. L'indicatrice *cens* est égal à 1 si l'événement est observé, 0 sinon.

Dans une première version de l'exemple nous choisissons de modéliser le risque de base par une constante spécifique à chaque intervalle de temps. Le suivi le plus long étant de 4 périodes, nous devons créer 4 indicatrices dont les coefficients transformés, pour repasser de $\hat{\beta}_{0t}$ à $\hat{h}_{ot}$, fourniront les estimations du risque de base. Le fichier de données correspondant est présenté dans la table 5.2.

Dans une seconde version nous supposons que β_{0t} varie linéairement avec les durées selon $\beta_{0t} = \alpha_0 + \alpha_1 t$. La table 5.3 présente la structure des données alors adaptée à l'estimation du modèle sous cette hypothèse.

Les explicatives x_1 et x_2 sont ici constantes dans le temps mais vous devez comprendre qu'il est possible d'avoir dans le fichier retraité des variables dont les valeurs se modifient au cours du temps. Ces variables peuvent être propres à chaque individu [7], mais elles pourraient aussi décrire l'environnement macro-économique de chaque période via par exemple un taux de chômage, d'inflation, etc...Dans ce cas, il faut naturellement être particulièrement attentif à la correspondance entre temps calendaire et temps d'événement pour être certain de faire correspondre la valeur d'une variable observée à une date quelconque à la durée qui lui correspond et cela pour chaque individu.

En pratique, avec SAS vous pouvez estimer la régression logistique avec la proc `logistic` en mettant la variable y_t en expliquée et en n'omettant pas l'option `descending` dans l'appel de la

7. Dans ce cas il est conseillé d'introduire ces variables avec des retards : lorsque les réalisations de l'événement et d'une explicative caractéristique d'un individu sont concomitants, il arrive souvent que la valeur de cette dernière devienne endogène, i.e. soit en fait expliquée par la survenue de l'événement plus qu'elle ne l'explique.

Id	y	x_1	x_2	D_1	D_2
1	0	x_{11}	x_{12}	1	1
1	0	x_{11}	x_{12}	1	2
1	1	x_{11}	x_{12}	1	1
2	0	x_{21}	x_{22}	1	2
2	0	x_{21}	x_{22}	1	3
3	0	x_{31}	x_{32}	1	1
3	0	x_{31}	x_{32}	1	2
3	0	x_{31}	x_{32}	1	3
3	1	x_{31}	x_{32}	1	4

TABLE 5.3 – Fichier d'exemple retraité pour l'estimation en temps discret avec un risque de base variable dans le temps selon $\beta_{0t} = \alpha_0 + \alpha_1 t$.

procédure afin que l'estimation porte sur $Prob[y_{it} = 1]$. En l'absence d'option précisant la fonction de lien dans la commande `model` alors par défaut `link=logit` et c'est donc la régression logistique qui est ajustée. Avec cette même procédure, vous pouvez également estimer la deuxième version du modèle, basée sur la transformée log-log, en faisant apparaître l'option `link=cloglog`.

Vous pouvez également utiliser la procédure `genmod` spécialisée dans l'estimation des modèles linéaires généralisés avec l'option `descending` dans son appel et en précisant dans la commande `model` la fonction de lien qui doit être utilisée, *i.e.* soit avec l'option `link=logit`, soit avec l'option `link=cloglog`. Vous retrouverez alors les résultats de la procédure `logistic` [8] .
Naturellement , avec les données de nos deux exemples précédents, il faudrait aussi pour des raisons que vous devez comprendre aisément spécifier l'option `noint`.

Si un des objectifs de votre étude est la construction de scores sur des individus non présents dans l'échantillon d'estimation alors on conseille l'emploi de la proc `logistic` qui autorise la sauvegarde d'un modèle estimé via l'option `outmodel=` mise dans l'appel de la procédure, sa relecture via l'option `inmodel=` dans un appel ultérieur de la proc, et son application sans ré-estimation sur un nouveau jeu de données, jeu constitué généralement des individus que l'on veut scorer, via la commande `score` [9].

5.2.1 Quels écarts-types utiliser ?

L'ajustement d'une régression logistique sur des observations telles que celles des tables 5.2 ou 5.3 est qualifié de *multiperiod logit* par Shumway [10]. Reprenant les travaux de Brown et Allison, cet

8. Attention cependant si vous utilisez la commande `class` : le codage des indicatrices alors créées n'est pas le même dans les deux procédures et donc leurs coefficients vont différer alors que les modèles estimés avec l'une ou avec l'autre sont en fait équivalents. Pour forcer le même comportement vous pouvez par exemple imposer l'option `param=glm` dans la commande `class` de la procédure `logistic`, cette dernière emploiera dans ce cas le codage mis en oeuvre par défaut par la procédure `genmod` et en conséquence les résultats devraient être strictement identiques

9. Pour mémoire, la procédure `PLM` peut également être employée pour construire ces scores. `PLM` est utilisable dès lors que la procédure qui a servi à l'estimation autorise la sauvegarde du modèle empirique ajusté via la commande `store`.

10. T. Shumway, "Forecasting Bankruptcy More Accurately : A Simple Hazard Model", *The Journal of Business* 74(1) :101-24, 2001.

auteur signale que les observations y_{it} concernant un même individu dans le multiperiod logit ne peuvent pas être indépendantes lorsque l'événement étudié ne peut pas se répéter. Par exemple, on ne peut pas connaître l'événement en t si on l'a subi en $t-1$, ou s'il survient en t c'est qu'il ne s'est pas réalisé en $t-1$. Pour tenir compte de cette non indépendance, Shumway propose de corriger les tests issus de la régression logistique. En effet, celle-ci utilise un nombre d'observations égal à $\sum_{i=1}^{n} t_i$ et donc, si $\hat{V}_\beta$ est la matrice de var-cov estimée des coefficients, les statistiques de Wald associées à $Ho : r\beta = r\beta_0$ seront de la forme $(1/\sum_{i=1}^{n} t_i)(r\beta - r\beta_0)^\top \hat{V}_\beta^{-1}(r\beta - r\beta_0)$. Comme on a réellement seulement n observations indépendantes [11], Shumway opère une pénalisation sur ces statistiques issues de la logistique de sorte à obtenir des valeurs égales à $(1/n)(r\beta - r\beta_0)^\top \hat{V}_\beta^{-1}(r\beta - r\beta_0)$. Cette correction qui semble raisonnable, et est d'ailleurs régulièrement utilisée dans nombre de travaux réalisés à la suite de cet auteur, entre en contradiction avec les recommandations d'Allison pour qui aucune correction ne doit être faite sur les écarts-types et les statistiques issues de la logistique. Le raisonnement tenu par Allison part des factorisations présentes dans les équations (5.6) et (5.7), factorisations qui vont mener à la fonction de log-vraisemblance d'un modèle à expliquée dichotomique. Soit pour mémoire :

$$\begin{aligned} Prob(T_i = t_i) &= h(t_i) \times [1 - h_i(1)][1 - h_i(2)] \dots [1 - h_i(t_i - 1)], \text{ et,} \\ S(t_i) &= [1 - h_i(1)][1 - h_i(2)] \dots [1 - h_i(t_i)] \end{aligned}$$

où $h(t_i)$ est la probabilité de subir l'événement en t_i sachant qu'il ne s'est pas réalisé auparavant. Dans ces équations, les produits des termes $h(t_i)$ et [1-$h(t_i)$] sont obtenus au moyen du théorème de Bayès et non pas en supposant qu'ils renvoient à des événements indépendants entre eux comme semble le suggérer leur lecture. En conséquence la log-vraisemblance à laquelle on parvient, donnée par (5.8), et qui est maximisée par la proc logistique, doit bien être ajustée sur la totalité des $\sum_{i=1}^{n} t_i$ observations et conduit à des estimateurs du maximum de vraisemblance vérifiant les propriétés usuelles de ces estimateurs. En d'autres termes, on donne raison à Allison : il n'est pas nécessaire de corriger les statistiques issues de cette estimation.

11. Dans l'exemple précédent, nous avons 3 individus supposés indépendants pour ajuster un modèle de survie, mais avec $t_1 = 3, t_2 = 2, t_3 = 4$, la régression logistique se fera sur 9 observations dont les 3 premières ne sont pas indépendantes entre elles, les deux suivantes non plus, de même que les quatre dernières.

Chapitre 6

Suppléments

6.1 Statistiques complémentaires ou alternatives aux ratios de risque

On comprend aisément la popularité des ratios de risque : ils résument en un seul chiffre les écarts de risque de différents individus. La publication de ces ratios est ainsi quasiment systématique dans les travaux empiriques estimant un modèle de Cox. Pour autant on ne doit pas ignorer leurs limites. Au delà d'une mauvaise spécification toujours possible du modèle estimé, par exemple due à l'omission d'explicatives pertinentes, il se peut que l'hypothèse centrale du modèle de Cox, à savoir la constance de la proportionnalité des risques, soit fausse [1]. Dans ce cas le ratio estimé est évidemment un estimateur biaisé du vrai ratio de risque. Un autre inconvénient est que le ratio renseigne seulement sur un risque relatif, il ne dit rien sur les niveaux absolus des risques concernés. De ce point de vue les courbes de survie qui renseignent sur la probabilité de survenue de l'événement étudié donnent plus d'information que le ratio de risque. Par ailleurs l'estimation du ratio dépend de la longueur de la fenêtre d'observation : deux études utilisant les mêmes individus mais des fenêtres d'observation inégales peuvent obtenir des ratios de risque différents.
En conséquence, d'autres statistiques ont été proposées pour pallier ces limites et/ou pour simplement donner des informations complémentaires sur les différences de survie d'individus ne possédant pas les mêmes caractéristiques. Nous ne présenterons ici que trois statistiques parmi les plus utilisées qui sont la durée de survie moyenne (*mean survival time*) et ses deux dérivées que sont la durée de survie moyenne restreinte (*restricted mean survival time*) et la durée de survie moyenne restreinte perdue (*restricted mean time loss*). Ces deux dernières durées sont évaluées conditionnellement à une durée de suivi des individus donnée justifiant le qualificatif de statistiques restreintes. Enfin nous mentionnerons la durée de survie médiane ('*median survival time*') et la survie conditionnelle.

6.1.1 Mean survival time, restricted mean survival time, restricted mean time loss

la durée de survie moyenne

Si t_k est le temps d'événement maximal et si T est une aléatoire supposée continue de densité $f(t)$, alors la durée de vie moyenne est donnée par :

1. Le test de l'hypothèse PH sera présenté ultérieurement

$$\mu = \int_0^{t_k} t f(t) dt, \tag{6.1}$$

et on montre aisément [2] que

$$\mu = \int_0^{t_k} S(t) dt, \tag{6.2}$$

où S(t) est la fonction de survie. La durée de survie moyenne est donc l'aire de la surface située sous la courbe de survie. Son estimateur naturel est :

$$\hat{\mu} = \int_0^{t_k} \hat{S}(t) dt, \tag{6.3}$$

où $\hat{S}(t)$ est l'estimateur de la survie, par exemple l'estimateur de Kaplan-Meier et dans ce cas $\hat{\mu}$ est évidemment un estimateur non paramétrique. Dans le cas le plus courant de censure à droite, une difficulté intervient avec ce mode de calcul lorsque la durée d'événement maximale correspond à une censure puisque dans ce cas l'aire en question n'est pas bornée à droite. En d'autres termes, l'estimateur donné par (6.3) n'est valide que lorsque la durée maximale est associée à un événement observé. Dans le cas contraire, cet estimateur sous-estime la survie moyenne.

restricted mean survival time, rmst

En raison de la difficulté précédente, à la durée moyenne de survie on préfère la durée moyenne de survie conditionnellement à une période de suivi des individus. Soit τ la longueur de cette période, la rmst est donc donnée par :

$$rmst(\tau) = \int_0^{\tau} S(t) dt, \tag{6.4}$$

qui sera estimée par

$$\widehat{rmst}(\tau) = \int_0^{\tau} \hat{S}(t) dt = \sum_{i=1}^{n^*} \hat{S}(t_{i-1}) \times (t_i - t_{i-1}) + \hat{S}(t_{n^*}) \times (\tau - t_{n^*}), \tag{6.5}$$

où n^* est le nombre de durées d'événement observées inferieur à τ. *rmst* est donc la durée d'attente moyenne pour que survienne un événement lorsque l'on se met sur la fenêtre d'observation $[0, \tau]$. De ce fait, cette statistique a une interprétation simple qui fait sens lorsqu'il s'agit de comparer les survies de différents groupes. En outre, on peut affiner les comparaisons en faisant les calculs de *rmst* pour diverses valeurs de τ et représenter des courbes de rmst.
Pour illustrer la démarche, nous allons utiliser les données de la table 6.1 relatives à 20 individus pour lesquels cens=1 si la durée est censurée à droite et cens=0 si l'événement est observé à la durée indiquée.

2. Je vous invite à retrouver vous-mêmes ce résultat.

time	2	2	3	4	4	5	6	7	8	8	8	8	10	10	12	12	13	14	14	15
cens	0	0	1	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	0

TABLE 6.1 – données d'exemple pour illustrer le calcul de rmst

La courbe de survie estimée sur ces données est présentée dans la figure 6.1 et permet l'obtention aisée des statistiques de rmst(τ) selon :

- sur $[0,2]$, $\widehat{rmst} = 2 \times 1 = 2$
- sur $[0,4]$, $\widehat{rmst} = 2 + 2 \times 0.9 = 3.8$
- sur $[0,7]$, $\widehat{rmst} = 3.8 + 3 \times 0.7941 = 6.1823$
- sur $[0,8]$, $\widehat{rmst} = 6.1823 + 1 \times 0.7330 = 6.9153$
- sur $[0,15]$, $\widehat{rmst} = 6.9153 + 7 \times 0.5498 = 10.7639$

Naturellement ce calcul peut être automatisé. Par exemple, si les données sont dans la table 'exemple', il suffit de faire :

```
* obtention et sauvegarde de l'estimateur de Kaplan-Meier;
proc lifetest data=exemple outs=km;
  time time*cens(1);
run;
* on ne conserve que les temps d'événements observés;
data km;
  set km;
  if _censor_ ne 1;
run;
* calcul de rmst pour chaque temps d'événement;
data rms;
  set km;
  rmst+(lag(survival))*(time-lag(time));
run;
proc print data=rms;
run;
proc sgplot data=rms;
  title "RMST(t)";
  series x=time y=rmst;
run;
```

On reproduit ci-après, la table 6.2 issue de la proc print présente dans cet ensemble de commandes qui confirme les résultats obtenus manuellement ainsi que la courbe de rmst créée par la `proc sgplot` dans la figure 6.2.

L'expression de l'estimateur de l'écart-type de $\widehat{rmst}$ est complexe. Vous pouvez la trouver [3] par exemple dans l'aide de la `proc lifetest` au moins dans SAS/STAT 15.1, version à partir de

3. Pour une démonstration, Royston P. and Parmar MKB, Restricted mean survival time : an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome, in *BMC Medical Research Methodology 2013, 13 :152*.

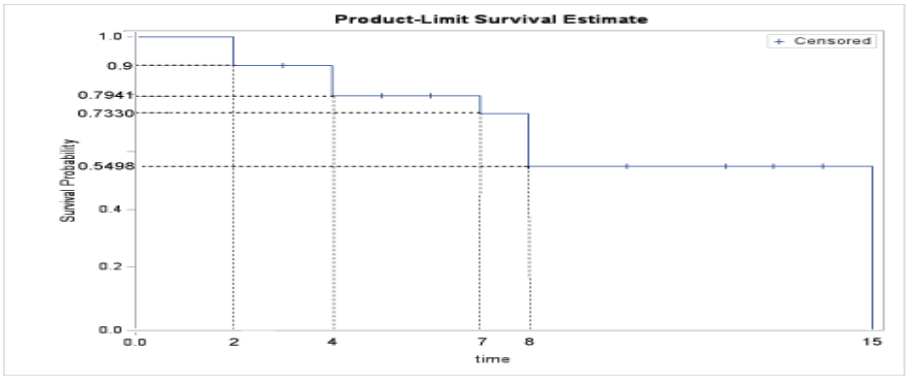


FIGURE 6.1 – calcul de RMST : estimation de S par Kaplan-Meier

Obs	time	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL	rmst
1	0	.	1.00000	1.00000	1.00000	0.0000
2	2	0	0.90000	0.65603	0.97401	2.0000
3	4	0	0.79412	0.53968	0.91745	3.8000
4	7	0	0.73303	0.47004	0.88006	6.1824
5	8	0	0.54977	0.29448	0.74621	6.9154
6	15	0	0.00000	.	.	10.7638

TABLE 6.2 – valeur de $\widehat{rmst}$ selon les longueurs de suivi des individus

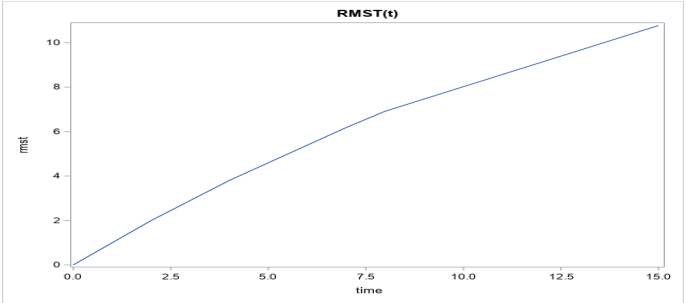


FIGURE 6.2 – courbe de RMST selon la longueur du suivi

RMST Analysis Information	
Tau	8

RMST Estimates	
Estimate	Standard Error
6.915385	0.4624

TABLE 6.3 – valeurs de rmst et de son écart-type obtenues avec `rmst(tau=8)`

laquelle on peut utiliser une option `rmst` : il est maintenant possible d'obtenir les valeurs de la statistique afférente à un suivi de longueur τ, de son écart-type et le graphe de la courbe de rmst. Il suffit pour cela de faire [4] :

```
proc lifetest data=... plots=(rmst) rmst(tau= τ);
```

ici, `plot=(rmst)` réclame l'affichage de la courbe de rmst, *i.e.* redonnerait la figure 6.2 que l'on vient de construire précédemment. Par ailleurs, en précisant une valeur pour τ, on obtient la valeur de la rmst sur la fenêtre de suivi $[0, \tau]$. Par exemple, toujours sur les données de notre exemple, l'emploi de `rmst`($\tau = 8$) renverra la table 6.3 et vous pouvez constater que la valeur affichée ici est identique à celle présentée dans la table 6.2.

Notez enfin que si l'on a deux groupes a et b, alors à τ donné :

$$\begin{aligned}\widehat{rmst}_a - \widehat{rmst}_b &= \int_0^\tau \hat{S}_a(t)dt - \int_0^\tau \hat{S}_b(t)dt \\ &= \int_0^\tau \left[\hat{S}_a(t) - \hat{S}_b(t)\right] dt\end{aligned}$$

la différence des *rmst* est simplement l'aire de la surface comprise entre les deux courbes de survie. Ainsi, si les échantillons a et b sont constitués d'individus indépendants, il est possible de tester l'égalité des deux durées de survie moyennes restreintes, via une z-transform :

$$z = \frac{\widehat{rmst}_a - \widehat{rmst}_b}{\sqrt{\hat{\sigma}^2_{rmst_a} + \hat{\sigma}^2_{rmst_b}}}$$

restricted mean time loss, rmtl

Alors que RMST indique la durée moyenne sans événement pour des individus suivis sur une période de temps τ, la statistique rmtl va donner la durée moyenne perdu par les mêmes individus au cours de la période de longueur τ. Les deux statistiques sont évidemment étroitement liées puisque formellement

$$rmtl = \tau - rmst = \int_0^\tau [1 - S(t)]dt \tag{6.6}$$

4. On peut obtenir ces statistiques sous *R* au moyen du package *survRM2*. Ceux qui sont intéressés peuvent consulter avec profit Hajime Uno, *Vignette for survRM2 package : Comparing two survival curves using the restricted mean survival time*, february 2017, Dana-Farber Cancer Institute.

et donc

$$rmtl + rmst = \widehat{rmtl} + \widehat{rmst} = \tau$$

La figure 6.3 illustre le calcul de ces statistiques sur les données de l'exemple précédent et pour $\tau = 8$. Si on suppose que les durées sont mesurées en années, les interprétations seraient les suivantes :

— avec $\tau = 8$ et $\widehat{rmst} = 6.92$ ans, on estime que si on suit ces individus pendant 8 ans, ils connaîtront en moyenne l'événement après environ sept ans .

— avec $\widehat{rmtl} = 1.08$ on estime que du fait de la réalisation de l'événement étudié, les individus perdent en moyenne une année sur les huit qui leurs sont allouées.

En pratique l'estimation de rmtl peut, à partir de la version SAS/STAT 15.1, s'effectuer via la proc `lifetest` selon une syntaxe similaire à celle vue pour l'estimation de rmst, *i.e.* l'exécution de :

```
proc lifetest data=... plots=(rmtl) rmtl(tau=8);
```

renverra le graphique de la courbe de rmtl et la valeur de la statistique pour $\tau = 8$.

test d'égalité des rmst et des rmtl entre groupes d'individus

On peut également tester l'égalité des rmst et rmtl entre divers groupes d'individus repérés par les modalités d'une variable spécifiée dans la commande `strata=` . Ainsi si la variable genre possède deux modalités repérant les hommes et les femmes, on pourra par exemple exécuter

```
proc lifetest rmst(tau=10) rmtl(tau=10) data=...;
time ...;
strata genre;
```

pour récupérer les valeurs des deux statistiques sur chaque sous-population pour une longueur de la période de suivi définie par $\tau = 10$, avec un test d'égalité des rmst et rmtl des hommes et des femmes, ici distribué selon un chi2 à 1 degré de liberté. Naturellement ce test peut être réalisé avec plus de deux modalités : Soit une variable *categories* possédant K modalités, l'hypothèse nulle devient :

$$H0 : rmst_1 = rmst_2 = \ldots = rmst_K, \text{ ou,}$$
$$H0 : rmtl_1 = rmtl_2 = \ldots = rmtl_K$$

l'hypothèse alternative étant qu'il existe au moins un couple pour lequel l'égalité n'est pas vérifiée. La statistique de test, qui obéira à un chi2 à $K - 1$ degrés de liberté est obtenue avec :

```
proc lifetest rmst(tau=10) rmtl(tau=10) data=...;
time ...;
strata categories;
```

Il s'agit bien évidemment d'un test joint et comme toujours, en cas de rejet on est souvent amené à rechercher le ou les couples responsables du rejet ce qui conduit à la réalisation de tests dans lesquels les statistiques sont prises deux à deux et réalisés

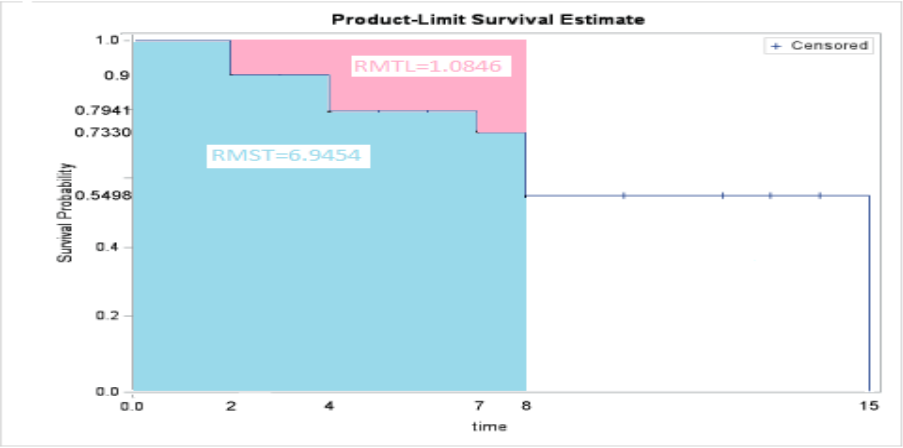


FIGURE 6.3 – Illustration : calcul de rmst(8) et de rmtl(8)

— soit pour tous les couples possibles, *i.e.*

$$H0 : rmst_i = rmst_j,\ i = 1, \ldots K,\ j = 1, \ldots K, i \neq j$$

au moyen de la syntaxe :
strata categories / dif=all ;

— soit en prenant une des modalités en référence et on examine alors les écarts des autres modalités à cette base. Par exemple, si on veut mettre $rmst_1$ en référence alors

$$H0 : rmst_i = rmst_1,\ i = 2 \ldots, K$$

et on exécutera :
strata categories / dif=control('$rmst_1$') ;

Comme une séquence de tests est réalisée dans les deux cas, on sait que pour tenter de stabiliser globalement le seuil de risque de première espèce de chacun au niveau désiré, il est nécessaire de réaliser un ajustement sur les seuils de risque de ces tests individuels. Cela peut se faire en mobilisant une correction, de type Sidak ou Bonferroni par exemple. Cette dernière sera ainsi réalisée au moyen de

strata categories / dif=control('$rmst_1$') adj=bon ;

6.1.2 Survie médiane

Il s'agit de la durée pour laquelle la moitié des individus ont connu l'événement étudié. On sait qu'il s'agit d'un estimateur plus robuste que la moyenne pour évaluer le centre d'une distribution. Pour cette raison, la survie médiane est souvent utilisée pour comparer des groupes d'individus.

Une façon simple de l'estimer est d'utiliser l'estimateur de la survie de Kaplan-Meier en tirant une horizontale passant par la probabilité de survie égale à 50% puis en redescendant sur l'échelle

des durées à partir du point d'intersection de cette horizontale et de la courbe de survie estimée. En procédant ainsi, avec les données utilisées dans notre exemple, on peut évaluer la durée médiane à 15 années. Notez cependant que deux configurations peuvent compliquer cette façon de faire :

— lorsque la courbe de survie n'est pas définie pour une probabilité de 50%. En général cela survient lorsque la proportion de données censurées à droite est élevée.
— lorsque l'horizontale à 50% se superpose à la courbe de survie estimée : ici on n'estime pas une durée médiane unique, mais une infinité de durées médianes. Dans notre exemple, il suffit d'imaginer qu'au lieu de la probabilité de 0.5498 nous lisions 0.5 pour s'apercevoir que la durée médiane serait alors n'importe où sur la plage [8 ans-15 ans].

Dans les deux cas la solution consiste simplement à travailler avec un autre fractile, par exemple 75% si un point d'intersection unique existe pour cette probabilité. D'ailleurs, même en l'absence de ces deux difficultés, on peut conseiller de réaliser des études de sensibilité, *i.e.* comparer les survies médianes de deux ou plusieurs groupes et examiner également les différences de survie correspondant aux premier et troisième fractiles par exemple afin d'améliorer la robustesse des conclusions énoncées.

6.1.3 Survie conditionnelle

La fonction de survie $S(u), u > 0$ nous indique la probabilité de connaître un événement après une durée u. Il peut être intéressant d'étudier l'évolution de la survie au-delà de cette durée u sachant que l'événement étudié ne s'est pas encore produit en u. Formellement, avec les notations habituelles, on s'intéresse alors à la probabilité conditionnelle : $Prob[T > u + t|T > u]$. Il vient immédiatement :

$$\begin{aligned} Prob[T > u + t|T > u] &= \frac{Prob[T > u + t, T > u]}{Prob[T > u]} \\ &= \frac{Prob[T > u + t]}{Prob[T > u]} \\ &= \frac{S(u + t)}{S(u)} \end{aligned} \tag{6.7}$$

Afin d'illustrer l'intérêt de cet indicateur, on reprend les données de notre exemple en supposant que l'événement d'intérêt est le décès suite à une certaine forme de cancer. Dans le cas de certains cancers, on pense que le risque de décès peut être élevé lorsque la maladie est diagnostiquée puis que ce risque diminue après que la personne ait été mise sous traitement. C'est un cas type où l'intérêt des probabilités conditionnelles peut apparaître. Dans l'exemple, nous estimons la survie à 8 ans après le diagnostic à un peu plus d'une chance sur deux (0.5498). L'estimation de la survie à 8 ans sachant que l'on a déjà survécu 4 ans est égale à

$$Prob[\text{décès après 8 ans} \mid \text{survécu déjà 4 ans}] = \frac{0.5498}{0.7941} = 0.69$$

Si vous considérez qu'il s'agit du cas d'une personne allant chez un assureur négocier un contrat, les conditions de celui-ci, et notamment le montant de la prime peuvent s'améliorer si l'on fait état des survies conditionnelles plutôt que des survies non conditionnelles. Cela a même été jugé suffisamment important pour que l'Institut national du cancer (Inca), Santé publique France, le

réseau des registres des cancers (réseau Francim) et le service de biostatistique des Hospices civils de Lyon (Rhône) décident de publier pour la première fois en 2018 des statistiques de probabilités conditionnelles.

6.1.4 Un exemple

Afin d'illustrer les points précédents, nous allons utiliser les données de Freireich déjà vues dans le premier chapitre lors de la présentation du test de log rank. Pour mémoire, elles donnent les temps de survie exprimés en semaines de patients leucémiques avec traitement 6-MP (groupe 1, 21 patients) et sans traitement (groupe 2, 21 patients). Nous avons simplement demandé l'exécution du programme suivant pour $\tau = 5$:

```
proc lifetest data=freireich plot=(s lls rmst rmtl)
  rmst(tau=5) rmtl(tau=5);
  time time*cens(1);
  strata groupe / test=(logrank wilcoxon);
run;
```

Afin d'étudier la sensibilité des conclusions à la longueur de la période de suivi, nous avons répété l'étude avec $\tau = 10$.

Les estimateurs des fonctions de survie sont représentés dans la figure 6.4. On peut déja remarquer que la survie estimée des patients du groupe 1 est, quelle que soit la durée considérée, supérieure à celle des patients du groupe 2. La figure 6.5 représente les transformées par la fonction log – log des survies estimées. Visuellement on n'a pas de raison de douter de leur parallélisme : l'hypothèse de risque proportionnel serait raisonnable, favorisant donc l'emploi de la statistique de log-rank pour le test d'égalité des survies. Cette statistique, ainsi que celle de Wilcoxon, est données dans la table 6.4. Les deux statistiques permettent de rejetter l'hypothèse d'égalité : le traitement de type 6-MP favorise significativement la survie des patients.

rmst et rmtl

Les courbes de restricted mean survival time obtenues sur chacun des groupes sont représentées dans la figure 6.6. Celle afférente au groupe 1 est toujours située au dessus de celle du groupe 2 : quelle que soit la durée du suivi des malades, la survenue du décès est en moyenne plus tardive chez les patients traités par 6-MP que chez les autres. C'est évidemment ce que dit aussi le graphique des rmtl de la figure 6.7 : le nombre de semaines perdues est toujours plus important pour les membres du groupe non traité. Vous devez avoir compris qu'il n'y a pas plus d'information dans l'un des graphiques que dans l'autre puisque quelle que soit la durée de suivi τ, on a $\tau = rmst(\tau) + rmtl(\tau)$.
Les valeurs estimées de rmst(5) et rmst(10) sont présentées dans la table 6.5. L'interprétation que l'on peut donner, par exemple avec $\tau = 10$, est que si on suit les deux groupes de patients pendant les dix semaines qui suivent le diagnostic, un décès est observé en moyenne après environ neuf semaines pour ceux du groupe 1 et après seulement six semaines et demie chez ceux du groupe 2. En d'autres termes, sur une fenêtre de suivi de dix semaines les personnes du groupe 1 vivent en moyenne 2 semaines 1/2 de plus que celles du groupe 2, et en moyenne presque une semaine

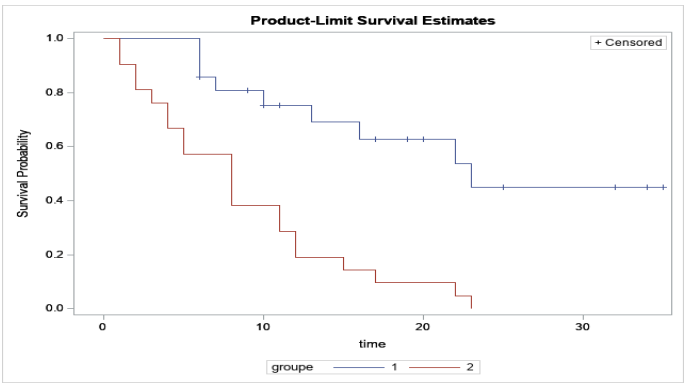


FIGURE 6.4 – Estimations des fonctions de survie, données de Freireich

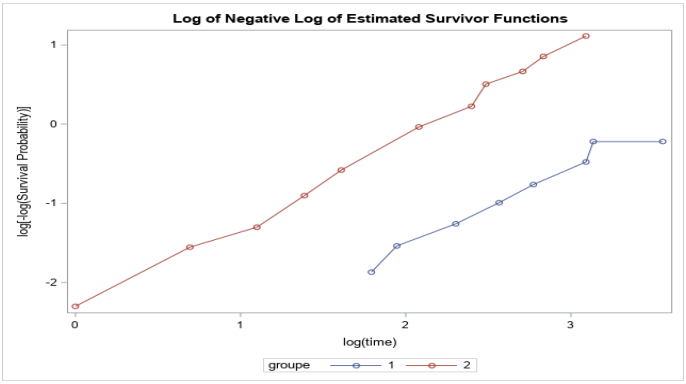


FIGURE 6.5 – Transformées log – log $\hat{S}$, données de Freireich

de plus sur un suivi de 5 semaines. Le test d'égalité signale que cet écart de durée de survie est significativement non nul conditionnellement à $\tau = 5$ semaines et à $\tau = 10$ semaines.
Pour illustration, on donne dans la table 6.6 les chiffres obtenus avec les mêmes valeurs de τ pour la statistique rmtl. On vous en laisse faire les commentaires et vous persuader que les informations données par ces tables 6.5 et 6.6 sont bien identiques.

Rappelez-vous également que le calcul explicite des écarts de rmst et/ou de rmtl entre les observations des différentes strates peut être obtenu en ajoutant l'option `dif=all` ou `dif=control('ref')` à la commande `strata`. Ainsi, en remplaçant la commande précédente par

```
strata groupe / test=(logrank wilcoxon) dif=all;
```

on obtient alors, en plus des résultats déja présentés, les tables intitulées *Restricted Mean Survival Time Comparisons* de la figure 6.7. Des affichages similaires, non reproduits ici, sont naturellement également donnés pour la statistique rmtl.

Cet exemple montre que ces statistiques rmst et rmtl ont une interprétation à la fois simple et parlante. Elles sont donc tout particulièrement utiles lorsque l'on veut présenter l'efficacité d'une action ou d'un traitement, ou plus généralement des différences de survie entre deux ou plusieurs groupes d'individus, à un public de non spécialistes. D'ailleurs, indépendamment de la validité ou non de l'hypothèse PH, plusieurs auteurs demandent aujourd'hui que leur publication soit aussi

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	16.7929	1	<.0001
Wilcoxon	13.4579	1	0.0002

TABLE 6.4 – Tests d'égalité des survies, données de Freireich

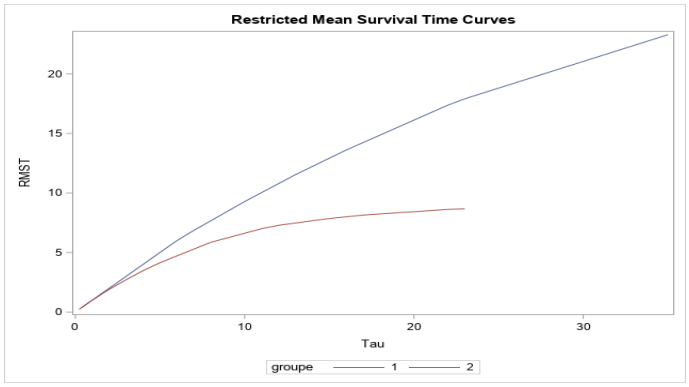


FIGURE 6.6 – Courbes des rmst, données de Freireich

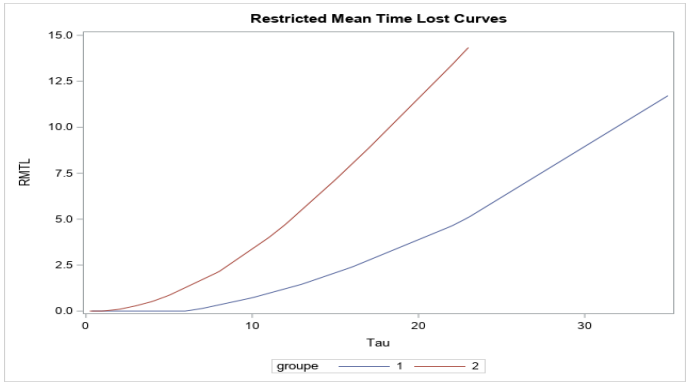


FIGURE 6.7 – Courbes des rmtl, données de Freireich

RMST Analysis Information	
Tau	5

RMST Estimates			
Stratum	groupe	Estimate	Standard Error
1	1	5	0.0000
2	2	4.142857	0.3033

RMST Test of Equality			
Source	Chi-Square	DF	Pr > ChiSq
Strata	7.9859	1	0.0047

RMST Analysis Information	
Tau	10

RMST Estimates			
Stratum	groupe	Estimate	Standard Error
1	1	9.277311	0.3268
2	2	6.619048	0.7330

RMST Test of Equality			
Source	Chi-Square	DF	Pr > ChiSq
Strata	10.9712	1	0.0009

TABLE 6.5 – Données de Freirech : rmst(5) et rmst(10) estimés sur les deux groupes avec test d'égalité.

RMTL Analysis Information	
Tau	5

RMTL Estimates			
Stratum	groupe	Estimate	Standard Error
1	1	0	0.0000
2	2	0.857143	0.3033

RMTL Test of Equality			
Source	Chi-Square	DF	Pr > ChiSq
Strata	7.9859	1	0.0047

RMTL Analysis Information	
Tau	10

RMTL Estimates			
Stratum	groupe	Estimate	Standard Error
1	1	0.722689	0.3268
2	2	3.380952	0.7330

RMTL Test of Equality			
Source	Chi-Square	DF	Pr > ChiSq
Strata	10.9712	1	0.0009

TABLE 6.6 – données de Freireich : rmtl(5) et rmtl(10) estimées sur deux groupes et test d'égalité, .

RMST Analysis Information	
Tau	5

Restricted Mean Survival Time Comparisons					
Stratum	Comparison	Difference	Standard Error	Chi-Square	Pr > ChiSq
1	2	0.857143	0.3033	7.9859	0.0047

RMST Analysis Information	
Tau	10

Restricted Mean Survival Time Comparisons					
Stratum	Comparison	Difference	Standard Error	Chi-Square	Pr > ChiSq
1	2	2.658263	0.8025	10.9712	0.0009

TABLE 6.7 – Tests de nullité des écarts des rmst entre strates, données de Freireich

%	groupe 1	groupe 2
75	.	12
50	23	8
25	13	4

TABLE 6.8 – Estimation des quartiles de la survie

systématique que celle des ratios de risque.

survie médiane

Dans la table 6.8 nous avons indiqué les durées de survie correspondant à la médiane ainsi qu'aux premier et troisième quartiles. Ces indicateurs, d'interprétation simple, peuvent également s'avérer très utiles pour faire ressortir les points forts d'une étude sur les durées de survie. Ici, on note que la moitié des patients du groupe 2 décèdent en moins de 8 semaines, contre 23 semaines pour ceux du groupe 1, qu'un individu sur 4 meurt avant 4 semaines pour le groupe non traité et avant 13 semaines pour le groupe traité au 6-MP. Enfin, si 75% des malades du groupe 2 décèdent en moins de 12 semaines, les trois-quarts des personnes ayant reçu le nouveau médicament sont toujours en vie au terme des 35 semaines de suivi de sorte qu'on ne peut pas estimer la durée associée à ce fractile. Ces chiffres font clairement ressortir l'intérêt du traitement 6-MP même auprès d'un non spécialiste des modèles de survie. Rappelez-vous aussi que la *proc lifetest* permet également de préciser des intervalles de confiance autour de la survie médiane et des survies à Q1 et à Q3.

6.2 L'ajustement de rmst en présence de variables explicatives

Dans la section précédente nous avons vu l'intérêt de la statistique *rmst* pour une présentation aisément compréhensible des résultats d'une étude des temps de survenu d'un événement ainsi que la simplicité de son calcul. Il faut cependant être conscient que ce dernier aspect est étroitement lié au cadre d'analyse qui était alors retenu et qui permettait d'évaluer la survie des individus au moyen de l'estimateur non paramétrique de Kaplan-Meier. Naturellement son estimation se complique lorsqu'on ne peut plus recourir à K-M, par exemple lorsqu'en prenant en compte des variables nominales, la constitution de sous-échantillons devient trop complexe et/ou ne laisse que des effectifs insuffisants dans certains d'entre eux. C'est également le cas lorsqu'on veut prendre en considération l'impact d'explicatives continues qu'on ne désire pas discrétiser. On sait que dans ces configurations une solution possible est l'ajustement d'un modèle de régression.

Précisément, des travaux récents permettent d'estimer une régression ayant *rmst(τ)* comme variable expliquée [5]. Un des avantages de ces régressions est qu'elles constituent une alternative à l'estimation d'un modèle de Cox ce qui est partculièrement utile lorsque l'hypothèse de risque

5. Andersen, P. K., Hansen, M. G., and Klein, J. P., Regression analysis of restricted mean survival time based on pseudo-observations, *Lifetime Data Analysis* 10, 335-350, 2004. Tian, L., Zhao, L., and Wei, L. J., Predicting the restricted mean event time with the subject's baseline covariates in survival analysis, *Biostatistics* 15, 222-233, 2014.

proportionnel est invalide. Pour réaliser ces régressions sous SAS il faut mettre en oeuvre la procédure `RMSTREG` disponible à partir de SAS/STAT 15.1 et à laquelle nous consacrerons cette section.

Il est déjà clair que les méthodes d'ajustements usuelles que vous connaissez ne fonctionneront généralement pas puisqu'en présence de censure à droite et de n individus ayant des caractéristiques différentes la variable expliquée $rmst_i, i = 1, \ldots$ n'est pas observée. Au moins deux solutions, présentes dans `rmstreg`, pourront cependant être employées : la régression sur pseudo-observations et la régression pondérée sur la probabilité inverse.

6.2.1 La régression sur pseudo-observations

Cette première méthode part d'une idée simple : puisque les vraies observations sont manquantes, on va les remplacer par des pseudo-observations. La difficulté est que le mécanisme de création doit assurer que les estimations de lois conditionnelles au moyen de ces pseudo-observations soient en un sens identiques à celles que l'on obtiendrait si on disposait des vraies observations.

Il se trouve que l'on peut satisfaire à cette exigence en ayant recours à la méthode du jacknife et aux équations d'estimation généralisées.

les pseudo-observations

Elles sont créées dans le cadre suivant : soit un ensemble d'aléatoires *i.i.d.* $\{T_1, T_2, \ldots, T_n\}$ et un paramètre θ du type $\theta = E[g(T)]$ où $g()$ est une fonction quelconque. On suppose que l'on sait calculer un estimateur sans biais de θ, soit $E[\hat{\theta}] = \theta$. Alors, par définition, la pseudo-observation correspondante à T_i est donnée par :

$$\hat{\theta}_i \equiv n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

où $\hat{\theta}^{-i}$ est calculé comme $\hat{\theta}$ mais sur l'échantillon duquel on a retiré la $i^{ème}$ observation. C'est la méthode du jacknife qui est à l'oeuvre ici.

— un premier exemple : les pseudo-observations associées à la fonction de survie. Soit la fonction indicatrice $\mathbf{1}(T > t)$ et $\theta = E[\mathbf{1}(T > t)] = Prob[T > t] = S(t)$. Soit $\hat{S}(t)$ l'estimateur de Kaplan-Meier de la survie. On sait qu'il est, sous certaines conditions, sans biais. Les pseudo-observations sont alors données par :

$$\hat{S}_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{-i}(t)$$

où $\hat{S}^{-i}(t)$ est l'estimateur de Kaplan-Meier de la survie dans l'échantillon privé du $i^{ème}$ individu.

— un second exemple : les pseudo-observations associées à la *rmst*. Celle-ci se définit par $rmst(\tau) = E[\min(T, \tau)]$. La fonction à considérer est donc simplement $\min(T, \tau)$ et ici $\theta = rmst(\tau)$. Par ailleurs on sait trouver un estimateur sans biais de $rmst(\tau)$: l'aire sous la courbe de survie estimée par Kaplan-Meier, $\hat{S}(t)$. Ainsi la pseudo-observation calculée pour le $i^{\text{ème}}$

individu sera :

$$\widehat{rmst}_i(\tau) = n\int_0^\tau \hat{S}(t)dt - (n-1)\int_0^\tau \hat{S}^{-i}(t)dt = \int_0^\tau \hat{S}_i(t)$$

$\hat{S}_i(t)$ étant la pseudo-observation de la survie définie dans l'exemple précédent.

— une remarque : supposons qu'il n'y ait pas de censure, dans ce cas θ serait simplement estimé par la moyenne empirique des transformées par $g()$ et donc :

$$\begin{cases} \hat{\theta} = \frac{1}{n}\sum_{j=1}^n g(T_j), \\ \hat{\theta}^{-i} = \frac{1}{n-1}\sum_{j=1, j\neq i}^n g(T_j) \end{cases} \quad \Rightarrow \hat{\theta}_i \equiv n\hat{\theta} - (n-1)\hat{\theta}^{-i} = g(T_i)$$

Dans cette situation la pseudo-observation n'est rien d'autre que la transformée par $g()$ de l'observation elle-même ainsi qu'on était en droit de l'attendre.

les équations d'estimation généralisées

L'idée est ensuite de remplacer les observations censurées par les pseudo-observations et d'ajuster une régression. C'est ce qu'ont proposé Andersen, Hansen, et Klein [6] dans le cadre d'un modèle linéaire généralisé d'écriture :

$$g(E[\,h(T_i)\,|\,X_i\,]) = \beta^\top X_i \tag{6.8}$$

où X_i est le vecteur des valeurs des explicatives pour le $i^{\text{ème}}$ individu et $g()$ est une fonction de lien que nous préciserons ci-après.

Pour l'estimation des β et de leurs covariances, ces mêmes auteurs ont proposé d'utiliser les équations d'estimation généralisées : avec cette technique, les coefficients des explicatives $\hat{\beta}$ sont obtenus en résolvant :

$$U(\beta) = \sum_{i=1}^n U_i(\beta) = \sum_{i=1}^n \left(\frac{\delta\, g^{-1}(\beta^\top X_i)}{\delta\beta}\right) V_i^{-1}\left(\widehat{rmst}_i - g^{-1}(\beta^\top X_i)\right) = 0 \tag{6.9}$$

où les $\widehat{rmst}_i$ sont les pseudo-observations et V_i leur variance. Un estimateur robuste de leur matrice de variance covariance, de type sandwich, est donné par :

$$\hat{\Sigma}_\beta = I(\hat{\beta})^{-1}\hat{var}(U(\beta))I(\hat{\beta})^{-1}$$

avec

$$I(\beta) = \sum_{i=1}^n \left(\frac{\delta\, g^{-1}(\beta^\top X_i)}{\delta\beta}\right)^\top V_i^{-1}\left(\frac{\delta\, g^{-1}(\beta^\top X_i)}{\delta\beta}\right)$$

$$\hat{var}(U(\beta)) = \sum_{i=1}^n U_i(\hat{\beta})U_i(\hat{\beta})^\top$$

Ce faisant, sous l'hypothèse que le mécanisme de censure est indépendant des explicatives, on obtient des estimateurs des coefficients qui sont asymptotiquement sans biais et gaussiens [7].

6. Andersen, P. K., Hansen, M. G., and Klein, J. P., Regression analysis of restricted mean survival time based on pseudo-observations, *Lifetime Data Analysis* 10, 335-350, 2004.

7. Graw, F., Gerds, T. A., and Schumacher, M., On pseudo-values for regression analysis in competing risks models, Lifetime Data Analysis 15, 241-255, 2009.

la fonction de lien

La fonction de lien $g()$ pourra prendre deux formes : soit celle de la fonction identité, soit celle de la fonction logarithmique. Nous pourrons donc ajuster :

1. $rmst_i(\tau) = \beta^\top X_i$, en spécifiant l'option `link=linear` ou `link=id` dans la commande `model`,
2. $\log(rmst_i(\tau)) = \beta^\top X_i$, en spécifiant l'option `link=log`

et la prévision de l'expliquée est donnée par $g^{-1}(\hat{\beta}^\top X_i)$.

Le choix de la fonction de lien n'est pas neutre pour l'interprétation des résultats : dans la première formulation, le coefficient d'une explicative mesure la variation de *rmst* que provoque l'augmentation d'une unité de cette explicative, les autres caractéristiques de l'individu étant inchangées. Dans la seconde, avec `link=log`, le même coefficient va évaluer le logarithme du rapport de *rmst* mesuré après l'augmentation d'un point de l'explicative à la *rmst* initiale, *i.e.* la variation relative de *rmst* générée par cette augmentation d'une unité.

Par ailleurs, pour le calcul des différences de *rmst*, on va disposer de commandes proches des commandes `hazardratio` et `contrast` de la procédure `phreg` :

— `estimate` : permet de réaliser des tests sur des combinaisons linéaires de coefficients selon une syntaxe qui rappelle celle de `contrast`. Attention cependant, la présence de son option `E` ne sert qu'à réclamer l'impression des pondérations de la combinaison linéaire testée ; pour prendre l'exponentielle du résultat, il faut mettre l'option `exp`. On dispose entre autres de l'option `at` permettant d'imposer des valeurs aux explicatives , de la construction d'intervalles de confiance qui peuvent être bilatéraux ou unilatéraux en précisant respectivement `cl`, `lower`, `upper` et naturellement de l'option `alpha=` .
— `lsmeans` permet le calcul de moyennes d'effets fixes. On dispose d'options déjà présentes dans `estimate` comme `at`, `dif` pour le calcul de différences de moyennes, `adjust=` en cas de comparaisons multiples,...
— `lsmestimate`, va permettre de réaliser des tests d'hypothèses sur les moyennes.

Notez que les commandes `lsmeans` et `lsmestimate` exigent l'emploi de `param=glm`, même si alors la désignation de la modalité en base peut être réalisée selon `la syntaxe habituelle nom variable(ref="modalité")`. Dans la section suivante nous donnons des exemples d'emplois de ces commandes.

Pour clore ce passage sur la fonction de lien, sachez que des règles de bon usage de l'une ou l'autre formulation ont été énoncées [8] : pour obtenir une inférence raisonnable sur les différences de *rmst* il faudrait avoir 10 événements observés ou plus par variable explicative, et 15 événements ou plus pour étudier les différences relatives de *rmst*.

8. Hansen, S. N., Andersen, P. K., and Parner, E. T., Events per variable for risks differences and relative risks using pseudo-obervations, Lifetime Data Analysis 20, 584-598, 2014.

Class Level Information				
Class	Value	Design Variables		
Group	AML-High Risk	1	0	0
	AML-Low Risk	0	1	0
	ALL	0	0	1

Convergence Status
Algorithm converged.

Type 3 Analysis of Effects			
Effect	DF	Chi-Square	Pr > ChiSq
Group	2	19.3581	<.0001

TABLE 6.9 – `RMSTREG, method=pv` : coefficients estimés, intervalles de confiance et significativité.

un exemple

Nous reprenons ici les données déjà utilisées de la table BMT. Il s'agissait d'étudier la durée de survie, appréciée en jours, de patients en fonction de la nature d'une tumeur [9], celle-ci étant classifiée en trois catégories intitulées respectivement : *ALL*, *AML-Low Risk* et *AML-High Risk*. Dans cet exemple nous n'utilisons que des variables catégorielles mais naturellement le modèle pourrait inclure des explicatives continues.

La première valeur à fixer est celle de la variable τ, *i.e.* la durée de temps sur laquelle nous allons évaluer la *rmst*. Ici, nous avons décidé de suivre les patients pendant 1100 jours, soit pratiquement 3 ans. Cette valeur de τ est imposée lors de l'appel de la procédure. Notez qu'il est possible de sauvegarder les pseudo-observations dans une table via l'option `outpv`=*nom de la table*. Une option `method=pv` peut être mise dans la commande model, sachant qu'en son absence, c'est la régression sur pseudo-observations qui est employée. Les commandes exécutées dans cet exemple sont les suivantes :

```
proc rmstreg data=bmt tau=1100 outpv=pseudo_obs;
format Group risk.;
class group(ref="ALL") / param=glm;
model t*status(0) = group / link=linear;
estimate "AML-High Risk vs AML-Low Risk" group 1 -1 0 / cl;
estimate "AML-High Risk vs ALL" group 1 0 -1 / cl;
estimate "AML-LOW Risk vs ALL" group 0 1 -1 / cl;
lsmeans group / cl ;
lsmestimate group 1 0 -1, 0 1 -1 / joint;
run;
```

La commande `model` possède la syntaxe usuelle, la seule spécificité ici concerne l'explicitation de la fonction de lien. Ici nous avons choisi la forme linéaire : les coefficients des explicatives s'interpréteront alors directement en termes de différence de niveaux de *rmst*.
Un premier ensemble de résultats obtenu est reproduit dans la table 6.9. Y sont rappelés les modalités crées par la commande ainsi que le test de significativité joint des coefficients des autres modalités et finalement une indication de l'état de sortie de l'algorithme d'estimation itératif. Ici on estime avoir convergé sur les estimateurs solutions de l'équation (6.9).

Les coefficients estimés sont renseignés dans la table 6.10. On y apprend que sur une fenêtre de 3 ans, la survie moyenne des patients ayant une tumeur qualifiée de *Low-Risk* est supérieure de

9. voir les données et le format employé à la page 28 ainsi que le graphique de la page 42.

Analysis of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	553.8729	69.8153	417.0375	690.7084	62.94	<.0001
Group	AML-High Risk	1	-138.560	93.6204	-322.053	44.9325	2.19	0.1389
Group	AML-Low Risk	1	225.4267	89.6665	49.6837	401.1698	6.32	0.0119
Group	ALL	0	0.0000					

TABLE 6.10 – `RMSTREG, method=pv` : coefficients estimés, intervalles de confiance et significativité.

Estimate							
Label	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
AML-High Risk vs AML-Low Risk	-363.99	84.0030	-4.33	<.0001	0.05	-528.63	-199.34

Estimate							
Label	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
AML-High Risk vs ALL	-138.56	93.6204	-1.48	0.1389	0.05	-322.05	44.9325

Estimate							
Label	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
AML-LOW Risk vs ALL	225.43	89.6665	2.51	0.0119	0.05	49.6837	401.17

TABLE 6.11 – `RMSTREG, method=pv` : Comparaison des survies moyennes associées à une fenêtre de suivi de 3 ans.

225 jours à celle des patients de la catégorie *All*, cet écart étant significativement non nul et que les personnes regroupées dans la modalité *High-Risk* survivent 138 jours de moins en moyenne que les individus référents, cet écart étant toutefois non significatif au seuil de 10%.

La syntaxe des commandes `estimate`, dont les résultats sont regroupés dans la table 6.11 doit vous être compréhensible, avec l'indication des pondérations à appliquer aux coefficients des différentes modalités de la variable *group* : il s'agit ici de comparer l'impact sur la rmst de l'appartenance à deux modalités différentes. Naturellement lorsque l'une de ces deux modalités correspond à la catégorie *All* on retrouve les mêmes informations que celles fournies par les résultats de l'estimation des coefficients dans la table 6.10. Par rapport à celle-ci, la seule nouveauté concerne les patients des modalités *Low-Risk* et *High-Risk* : on estime que sur un suivi de 3 ans, les premiers survivent en moyenne une année de plus que les seconds, cet écart étant significatif aux seuils de risque usuels.

L'estimation de la durée de survie moyenne évaluée sur une période de 3 ans, et non plus la différence de survie donnée par `estimate`, est fournie par la commande `lsmeans` qui est mobilisée ici avec son option `cl` réclamant donc aussi les bornes d'un intervalle de confiance à 95%. En sortie on obtient la table 6.12 dont la lecture ne doit pas soulever de difficultés. Par exemple, la survie moyenne des patients qualifiés de *Low-Risk* suivis pendant trois ans est estimée à 779 jours, soit un peu plus de deux années, l'intervalle de confiance à 95% étant de [669 jours-890 jours]. Bien évidemment, les valeurs affichées dans cette table sont compatibles avec les écarts de survie discutés précédemment.

Notez encore que l'on peut faire apparaître plusieurs commandes `lsmeans` dans le même programme. Par exemple, en plus de celle venant d'être présentée, nous aurions pu demander

Group Least Squares Means							
Group	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
AML-High Risk	415.31	62.3747	6.66	<.0001	0.05	293.06	537.57
AML-Low Risk	779.30	56.2663	13.85	<.0001	0.05	669.02	889.58
ALL	553.87	69.8153	7.93	<.0001	0.05	417.04	690.71

TABLE 6.12 – `RMSTREG, method=pv` : survies moyennes associées à une fenêtre de suivi de 3 ans.

Differences of Group Least Squares Means Adjustment for Multiple Comparisons: Bonferroni						
Group	_Group	Estimate	Standard Error	z Value	Pr > |z|	Adj P
AML-High Risk	AML-Low Risk	-363.99	84.0030	-4.33	<.0001	<.0001
AML-High Risk	ALL	-138.56	93.6204	-1.48	0.1389	0.4166
AML-Low Risk	ALL	225.43	89.6665	2.51	0.0119	0.0358

TABLE 6.13 – `RMSTREG, method=pv` : Différences de survies moyennes avec ajustement de Bonferroni.

l'exécution de :

```
lsmeans group / diff adjust=bon;
```

Celle-ci réclame des tests de significativité des différences de moyennes de rmst entre les différentes modalités de la variable *group*, avec ajustement de Bonferroni du seuil de risque pour comparaisons multiples. La table 6.13 est alors produite. Naturellement, à l'exception de cet ajustement du seuil, on retrouve des résultats déjà vus avec la commande `estimate`.

Enfin, la commande `lsmestimate` employée ici réclame un test joint de significativité des modalités *All-Risk* et *Low-Risk* : les pondérations utilisées réclament de part et d'autre du séparateur "," la comparaison de chacune d'elles avec la catégorie *All*, et l'option `joint` demande évidemment la construction d'un test joint. Les sorties obtenues sont présentées dans la table 6.14 où l'on voit que les deux premières catégories ont des coefficients, et donc des survies moyennes, significativement différents de la troisième à la fois individuellement et globalement. Fort heureusement, on retrouve évidemment des résultats déjà affichés, par exemple dans la figure 6.13 pour les tests individuels et dans la figure 6.9 pour le joint.

Least Squares Means Estimates					
Effect	Label	Estimate	Standard Error	z Value	Pr > |z|
Group	Row 1	-138.56	93.6204	-1.48	0.1389
Group	Row 2	225.43	89.6665	2.51	0.0119

Chi-Square Test for Least Squares Means Estimates			
Effect	Num DF	Chi-Square	Pr > ChiSq
Group	2	19.36	<.0001

TABLE 6.14 – `RMSTREG, method=pv` : Commande `lsmestimate` test d'hypothèses jointes.

6.2.2 La régression pondérée par les probabilités inverses

Une autre procédure d'estimation de l'équation de base (6.8) a été proposée par Tian, Zhao et Wei [10]. Pour estimer les coefficients $\hat{\beta}$ ces auteurs vont utiliser une régression de type moindres carrés pondérés faisant appel à la distribution estimée des temps de censure. Plus précisément, soit T et C les temps d'événement et de censure, alors les données observées sont $min(T, C)$, l'indicatrice $\delta = 1$ si $T \leq C$ (l'événement est observé) et 0 sinon, ainsi que les caractéristiques de l'individu, X. Soit τ une longueur de suivi des individus, on note Y le temps d'événement contraint par τ, *i.e.* $Y = \min(T, \tau)$. On sait que l'espérance de $rmst = E[Y]$. Par ailleurs si l'estimateur de KM de la survie est construit sur les valeurs $(min(T, C), \delta)$, on peut également calculer l'estimateur KM afférent aux temps de censure sur les valeurs $(min(T, C), 1 - \delta)$, i.e. la probabilité d'une censure au-delà d'une durée t. Soit $\hat{G}$ ce dernier estimateur et une autre indicatrice $\tilde{\delta} = 1$ si Y est observé et 0 sinon (l'individu a été effectivement suivi sur la durée τ et l'événement s'est réalisé, *i.e.*$Y = T$, ou ne s'est pas encore réalisé, *i.e.* $Y = \tau$, pendant ce suivi). Les coefficients des explicatives dans l'équation de rmst sont solutions de :

$$U(\beta) = n^{-1} \sum_{i=1}^{n} \omega_i X_i \left[Y_i - g^{-1}(X_i^\top \beta) \right] = 0 \tag{6.10}$$

avec la pondération $\omega_i = \frac{\tilde{\delta}_i}{\hat{G}(Y_i)}$: chaque individu pour lequel Y est observé se voit affecté d'un poids égal à l'inverse de sa probabilité d'être censuré. Il est encore possible d'obtenir un estimateur robuste de leur matrice de variance-covariance [11]. Notez encore que selon Tian, Zhao et Wei [12] on peut montrer que les estimateurs des pseudo-observations et ceux des régressions pondérées devraient généralement converger vers les mêmes valeurs si le modèle estimé est correctement spécifié.
Avant d'illustrer cette procédure d'estimation, un dernier point peut encore être discuté : dans le calcul des poids, l'estimation KM de la survie sur les temps de censure, $\hat{G}(Y_i)$ a été présenté ci-dessus en supposant l'homogénéité de la distribution des temps de censure sur la totalité des individus. On peut imaginer, s'il existe plusieurs strates d'individus, que cette hypothèse soit fausse. Dans ce cas, il est possible d'estimer non pas une survie sur la totalité de l'échantillon, mais une fonction $\hat{G}(Y_i)$ spécifique à chacune des strates, fonction qui naturellement sera utilisée sur les individus qui composent la strate en question. Comme nous allons le voir, il est aisé de construire un test de Log-rank ou de Wilcoxon d'égalité des distributions des censures entre sous-échantillons et donc de choisir d'estimer une seule ou plusieurs fonctions $\hat{G}(Y_i)$.

un exemple

Pour cet exemple de régression pondérée sur les probabilités inverses, nous reprenons les données utilisées qui viennent d'être employées pour illustrer la régression sur pseudo-observations. La mise en oeuvre de la régression pondérée s'effectue via la commande `model` et son option `method=ipcw`. Par défaut on rappelle que `method=pv`. La commande `model` possède donc maintenant la syntaxe suivante :

10. Tian, L., Zhao, L., and Wei, L. J., Predicting the restricted mean event time with the subject's baseline covariates in survival analysis, Biostatistics 15, 222-233, 2014.

11. Voir par exemple l'aide de la procédure `rmstreg` pour le détail des formules.

12. Tian, L., Zhao, L., and Wei, L. J., On the Restricted Mean Event Time in Survival Analysis, Harvard University Biostatistics Working Paper Series 156, 2013.

Analysis of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	550.7621	72.7474	408.1798	693.3445	57.32	<.0001
Group	AML-High Risk	1	-119.638	100.2510	-316.126	76.8508	1.42	0.2327
Group	AML-Low Risk	1	222.3733	93.5651	38.9890	405.7576	5.65	0.0175
Group	ALL	0	0.0000					

TABLE 6.15 – `RMSTREG, method=ipcw` : coefficients estimés, intervalles de confiance et significativité.

Group Least Squares Means							
Group	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
AML-High Risk	431.12	67.2435	6.41	<.0001	0.05	299.33	562.92
AML-Low Risk	773.14	56.7298	13.63	<.0001	0.05	661.95	884.32
ALL	550.76	72.7474	7.57	<.0001	0.05	408.18	693.34

TABLE 6.16 – `RMSTREG, method=ipcw` : survies moyennes associées à une fenêtre de suivi de 3 ans.

```
model t*status(0) = group / link=linear method=ipcw;
```

Une fois cette correction effectuée, et évidemment après avoir supprimé l'option `outpv=` dans l'appel de la procédure, toutes les commandes du précédent exemple fonctionnent. Avec ces données, les résultats des deux méthodes d'estimation sont proches, et conduisent qualitativement aux mêmes conclusions. Nous n'indiquerons ici dans la table 6.15 que les coefficients estimés, qui peuvent être comparés avec ceux de la table 6.10, et les estimations des rmst issues de la commande estimate de la table 6.16 que l'on peut rapprocher de celles affichées dans la table 6.12.

Le dernier point abordé dans cet exemple concerne l'hypothèse d'homogénéité des individus : on rappelle que les probabilités inverses sont issues d'une estimation de type KM sur les temps de censure observés dans l'échantillon complet. Afin de tester cette hypothèse de distribution identique des censures sur les strates afférentes aux 3 modalités de la variable *group*, on peut mettre en oeuvre un test d'égalité des courbes de survie. Il faut simplement renverser le statut des observations censurées et non censurées. Ainsi, dans notre exemple, la variable *status* vaut 0 lorsque le temps d'événement est censuré, d'où la syntaxe de la commande `model t*status(0) = ...` dans les exemples précédents. Le test qu'il nous faut mener maintenant considère que l'événement est la censure, il sera donc réalisé via la proc `lifetest` au moyen des instructions qui suivent :

```
proc lifetest data=bmt;
format Group risk.;
strata group;
time t*status(1);
run;
```

Notez bien l'emploi de `status(1)` qui va donc conduire à la comparaison entre les trois sous-échantillons des fonctions de survies évaluées sur les temps de censure. Ici nous obtenons les résultats présentés dans la table 6.17. L'hypothèse nulle est rejetée, ce qui invalide les calculs précédents. La solution est évidemment d'estimer les probabilités inverses à partir de courbes de survie estimées au sein des sous-échantillons et non pas sur leur mélange pour utiliser des

Testing Homogeneity of Survival Curves for T over Strata

Rank Statistics		
Group	Log-Rank	Wilcoxon
ALL	7.4459	298.00
AML-High Risk	-5.2622	-144.00
AML-Low Risk	-2.1837	-154.00

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	11.3481	2	0.0034
Wilcoxon	9.1030	2	0.0106
-2Log(LR)	0.8661	2	0.6485

TABLE 6.17 – Test d'identité des distributions des censures sur les strates de la variable group.

Analysis of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	552.8769	70.7980	414.1154	691.6384	60.98	<.0001
Group	AML-High Risk	1	-137.010	94.2156	-321.669	47.6489	2.11	0.1459
Group	AML-Low Risk	1	225.6872	90.1993	48.8997	402.4746	6.26	0.0123
Group	ALL	0	0.0000					

TABLE 6.18 – `RMSTREG, method=ipcw`, pondérations spécifiques aux modalités : coefficients estimés.

pondérations spécifiques à chacune des catégories de patients . Pour cela, il suffit d'employer la sous-option `strata=` de l'option `method=` . On pourra ainsi exécuter :

```
model t*status(0) = group / link=linear method=ipcw(strata=group);
```

Les autres commandes s'utilisent sans changement.

Avec nos données, la prise en compte de distributions de censure différentes n'affecte pratiquement pas les résultats des estimations comme le montre les tables 6.18 et 6.19 précisant les estimateurs des coefficients et des rmst alors obtenus : vous pouvez vérifier qu'ils diffèrent peu de ceux affichés dans les tables correspondantes vues précédemment.

6.3 Estimation non paramétrique avec censure par intervalle

Nous avons déjà présenté les divers cas de censures et on sait que l'approche paramétrique permet de les prendre en compte aisément. En revanche, les procs `lifetest` et `phreg` n'autorisent que des censures à droite. Vous pouvez toutefois trouver dans SAS deux procédures plus récentes, `iclifetest` et `icphreg`, qui vont permettre, sur données censurées à gauche et par intervalle,

Group Least Squares Means							
Group	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
AML-High Risk	415.87	62.1629	6.69	<.0001	0.05	294.03	537.70
AML-Low Risk	778.56	55.8889	13.93	<.0001	0.05	669.02	888.10
ALL	552.88	70.7980	7.81	<.0001	0.05	414.12	691.64

TABLE 6.19 – `RMSTREG, method=ipcw`, pondérations spécifiques aux modalités : survies moyennes

l'estimation non paramétrique de la fonction de survie d'une part et du modèle de Cox d'autre part.
Si elles sont dans leurs choix d'options et dans leurs sorties assez proches des procs `lifetest` et `phreg`, les algorithmes mis en oeuvre sont cependant sensiblement différents. Nous en ferons donc une brève présentation.

Tout d'abord, un peu de terminologie :

— censure par intervalle de type 1 : On a observé chaque individu i à une durée unique o_i à laquelle on sait seulement si l'événement étudié s'est ou non produit. En conséquence les intervalles concernant tout individu contiennent soit zéro dans le cas d'une censure à gauche, soit l'infini, dans le cas d'une censure à droite, *i.e.* si t_i est la durée de survenue de l'événement, ils sont de la forme $[0, o_i]$ si $t_i \leq d_i$ et $[o_i, \infty[$ sinon.

— censure par intervalle de type 2 : On dispose pour chaque individu de deux aléatoires, B_i et H_i dont les réalisations vont être les bornes d'un intervalle de censure et on sait seulement que la durée de survenue de l'événement étudié, t_i est telle que $b_i <= t_i <= h_i$. On retrouve évidemment les cas connus :
  — si $b_i = 0 \rightarrow$ censure à gauche,
  — si $h_i = \infty \rightarrow$ censure à droite,
  — si $b_i = h_i \rightarrow$ pas de censure.
  — si $(b_i = 0 \ \& \ h_i > 0)$ ou si $(b_i > 0 \ \& \ h_i = \infty) \ \rightarrow$ censure de type 1.

6.3.1 Les intervalles de Turnbull

Un des premiers algorithmes d'estimation non paramétrique de la survie sur données censurées par intervalle est dû à Turnbull [13].
Vous devez comprendre que lorsque les censures afférentes aux individus constituant l'échantillon ne se recoupent pas, alors les algorithmes écrits pour une censure à droite peuvent être utilisés presque sans modification. Il suffit de choisir une règle d'imputation pour le temps d'événement, par exemple le centre de l'intervalle de censure, et on est alors en mesure de comptabiliser le nombre d'individus à risque pour chaque temps d'événement et donc finalement de calculer l'estimateur de Kaplan-Meier.
Une difficulté intervient lorsque les intervalles de censures des individus ont des intersections non nulles ce qui est naturellement la situation la plus courante. Dans ce cas, Turnbull montre que l'estimateur de la survie n'est pas défini sur des intervalles particuliers dits intervalles de Turnbull. Ces derniers sont tels que leurs bornes basses et hautes font parties des bornes basses ou hautes des intervalles de censure initiaux mais également tels qu'ils ne contiennent aucune autre des bornes initiales. Plus précisément, soit $IC = \{ (b_i, h_i) \}, i = 1, \ldots n$ l'ensemble des intervalles de censure afférent à n individus, soit :

$L = \{b_{(l)}, b_{(2)}, b_{(n)}\}$ l'ensemble des bornes basses classées par ordre croissant et,
$H = \{b_{(l)}, b_{(2)}, b_{(n)}\}$ l'ensemble des bornes hautes également classées par ordre croissant,

alors les intervalles de Turnbull sont les intervalles disjoints tels que leur borne basse appartient à L, leur borne haute appartient à H et ne contiennent eux-mêmes aucune autre borne présente dans

13. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data, *Journal of the Royal Statistical Society. Series B* (Methodological) Vol. 38, No. 3 (1976), pp. 290-295.

i	b	h
1	0	2
2	5	7
3	4	6
4	8	10
5	6	9

TABLE 6.20 – exemple de 5 intervalles de censure

L ou H. Par exemple si on a observé les intervalles de censure de 5 individus précisés dans la table 6.20, alors

$$L = \{0, 4, 5, 6, 8\} \text{ et,}$$
$$H = \{2, 6, 7, 9, 10\}.$$

Les intervalles de Turnbull seront :]0,2],]5,6],]6,7] et]8,9]

6.3.2 L'estimation de la survie

1. Plusieurs estimateurs ont été proposés. Parmi ceux-ci, l'un des plus anciens dérive des travaux de Peto [14] et B. W. Turnbull [15]. Ce dernier montre d'une part qu'un estimateur du maximum de vraisemblance de la fonction de répartition ne peut pas être croissant en-dehors des intervalles de Turnbull et d'autre part que la vraisemblance dépend de l'accroissement de la répartition sur ces intervalles mais pas du chemin par lequel $F()$ passe du plancher au plafond [16]. La conséquence de ces deux résultats est que l'estimateur du maximum de vraisemblance de la répartition est plat entre les intervalles de Turnbull et indéterminé sur ces intervalles. Il en va alors naturellement de même pour l'estimateur du maximum de vraisemblance de la survie. L'algorithme de Turnbull permet de calculer un estimateur du maximum de vraisemblance non paramétrique de la survie en présence de données censurées par intervalle sous certaines hypothèses. Parmi celles-ci on va retrouver l'hypothèse de censure non informative : comme on l'a déjà vu, en présence d'un mécanisme de censure aléatoire et d'une durée de survenue également aléatoire, on peut simplifier l'écriture de la vraisemblance en supposant que la loi jointe afférente aux intervalles de censure ne contient pas d'information sur la distribution de durées de survenue. En effet, pour une individu i la vraisemblance fait a priori intervenir la loi jointe de l'aléatoire T_i dont la réalisation sera la durée effectivement observée, ou censurée, t_i et des deux aléatoires dont les réalisations sont les bornes des intervalles de censure, b_i et h_i. Avec l'hypothèse de censure non informative, la vraisemblance va seulement être proportionnelle à la probabilité que T_i appartienne à $[b_i, h_i]$. Sur un ensemble d'individus indépendants la vraisemblance à considérer va donc s'écrire :

$$\mathcal{L} = \Pi_{i=1}^{n} Pr\left[T_i \in [b_i, h_i]\right] = \Pi_{i=1}^{n} [S(b_i) - S(h_i)] = \Pi_{i=1}^{n} [F(h_i) - F(b_i]$$

14. Peto, R., Experimental survival curves for interval-censored data, *Applied Statistics*, 22, 86-91, 1973.
15. Turnbull, B. W., The empirical distribution function with arbitrarily grouped,censored and truncated data, *Journal of the Royal Statistical Society*, 38, 290-95, 1976.
16. Une image de ce résultat est que sur un intervalle de Turnbull seule compte la hauteur de la marche, *i.e.* l'accroissement de la répartition, pas la largeur de la marche.

L'estimateur du maximum de vraisemblance de la survie ne peut donc décroître que sur les intervalles de Turnbull et la seule valeur à estimer est l'ampleur de la baisse. Si on note $]p_1, q_1],]p_2, q_2], \ldots]p_m, q_m]$ les bornes des intervalles de Turnbull, alors la probabilité que la durée d'événement appartienne à un intervalle de censure peut être réécrite sur les probabilités qu'elle appartienne aux intervalles de Turnbull. Soit I_{ij} une indicatrice valant 1 si le $j^{\text{ème}}$ intervalle de Turnbull est contenu dans l'intervalle de censure du $i^{\text{ème}}$ individu :

$$\mathbf{1}_{ij} = \begin{cases} 1 \text{ si }]p_j, q_j] \in [b_i, h_i], \\ 0 \text{ sinon.} \end{cases}$$

En conséquence, en notant $\omega_j = Pr[p_j <= T_i <= q_j]$, la fonction de vraisemblance précédente s'écrit sur ces paramètres ω_j comme :

$$\mathcal{L}(\omega) = \prod_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}_{ij}\, \omega_j$$

Peto a proposé d'employer un algorithme de maximisation usuel, du type Newton-Raphson, dans lequel on impose les contraintes $\hat{\omega}_j \geq 0$ et $\sum_{j=1}^{m} \hat{\omega}_j = 1$ mais cet algorithme rencontre des difficultés de convergence lorsque le nombre de paramètres à estimer, *i.e.* le nombre d'intervalles de censure, augmente. Turnbull a toutefois montré que les solutions du problème précédent pouvaient être obtenues en résolvant le système d'équations simultanées suivant :

$$\omega_j = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}(\omega_1, \omega_2, \ldots, \omega_m) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}_{ij}\omega_j}{\sum_{j=1}^{m} \mathbf{1}_{ij}\omega_j}, j = 1, 2, \ldots, m$$

où $\mu_{ij}()$ est la probabilité espérée pour que la durée de survenue de l'événement pour l'individu i appartienne au $j^{\text{ème}}$ intervalle de Turnbull.

C'est un algorithme espérance-maximisation (EM) qui est à l'oeuvre : pour des valeurs initiales des paramètres ω on calcule la valeur de l'espérance, μ_{ij} (étape E), puis on met à jour les ω(étape M) avant de retourner à l'étape E. Les itérations s'arrêtent soit lorsque les révisions observées sur les paramètres, soit lorsque la variation de la vraisemblance sont considérées comme quasi nulles.

Soit $\hat{\omega}_j$, $j = 1, \ldots m$ les solutions obtenues, l'estimation de la fonction de survie est finalement donnée par :

$$\hat{S}(t) = \begin{cases} 1 \text{ si } t < p_1, \\ \hat{\omega}_2 + \hat{\omega}_3 + \ldots + \hat{\omega}_m \text{ si } q_1 \leq t < p_2 \\ \hat{\omega}_3 + \hat{\omega}_4 + \ldots + \hat{\omega}_m \text{ si } q_2 \leq t < p_3 \\ \vdots \\ \hat{\omega}_m \text{ si } q_{m-1} \leq t < p_m \\ 0 \text{ si } t \geq q_m \end{cases} \tag{6.11}$$

et est indéterminée si $p_j \leq t \leq q_j$ pour $j = 1, 2, \ldots m$.

Un défaut de cet algorithme est qu'il n'est pas certain de mener à un maximum global. Pour assurer la convergence vers un estimateur du maximum de vraisemblance, on a montré que les multiplicateurs de Lagrange associés aux $\hat{\omega}_j$ nuls devaient être positifs ou nuls. Les valeurs de ces multiplicateurs sont d'ailleurs affichés conjointement avec les estimations de la survie dans la table intitulée `Nonparametric Survival Estimates`.

2. l'algorithme de Groeneboom and Wellner [17] connu sous l'appellation *iterative convex minorant algorithm* (ICM) que nous ne détaillerons pas ici. Il maximise la logvraisemblance en mettant en oeuvre un algorithme de Newton-Raphson modifié et est plus rapide que l'algorithme de Turnbull.
3. l'algorithme de Wellner et Zhan, dénommé algorithme EMICM combine les deux précédents en basculant de l'un à l'autre au cours de ses itérations. Il a l'avantage de fournir une solution qui converge vers l'estimateur du maximum de vraisemblance non paramétrique si celui-ci est unique. C'est d'ailleurs EMICM qui est utilisé par défaut dans la procédure `ICLIFETEST`.

Dans l'appel de la procédure on trouvera donc une option `method=`*`sigle`*, où *`sigle`* prend l'un des intitulés suivants : `TURNBULL, ICM, EMICM`, sachant encore que l'algorithme de Turnbull peut être appelé indifféremment par `method=TURNBULL` ou `method=EM`.

6.3.3 Intervalles de confiance ponctuels sur la survie

Étant donné la construction de l'estimateur de la survie explicitée dans le système d'équations (6.11) on pourrait penser construire des intervalles de confiance autour de la survie estimée à partir de la matrice des variances-covariances estimées des $\hat{\omega}$, celle-ci étant obtenue via la matrice des dérivées secondes de la logvraisemblance. Ici cette théorie ne peut s'appliquer car le nombre de paramètres à estimer n'est pas constant mais augmente avec n. La procédure propose alors deux solutions pour estimer ces variances :

— la première méthode utilise des échantillons bootstrappés sur chacun desquels elle estime une fonction de survie pour finalement estimer une variance empirique. Ainsi, si on dispose de B échantillons bootstraps alors la variance de $\hat{S}(t)$ sera évaluée [18] par

$$var(\hat{S}(t)) = \frac{1}{B-1} \sum_{b=1}^{B} [\hat{S}_b(t) - \bar{S}_b(t)]^2, \text{ où}$$

$$\bar{S}_b(t) = \frac{1}{B} \sum_{b=1}^{B} \hat{S}_b(t)$$

17. Groeneboom, P., and Wellner, J. A. (1992). Information Bounds and Nonparametric Maximum Likelihood Estimation. Basel : Birkhäuser, 1992

18. Un ajustement est toutefois effectué sur les estimateurs de la survie des échantillons bootstrappés. Sans cela, comme à chaque échantillon correspond une ensemble d'intervalles de Turnbull qui lui est propre, on court le risque d'avoir des durées pour lesquelles la survie n'est pas définie.

— la seconde, dite d'imputations multiples, se contente de remplacer les intervalles censurés de la table de données par une durée d'événement unique prise au hasard dans $]b_i, h_i]$. Une estimation de Kaplan-Meier de $S(t)$ est ensuite réalisée. La procédure est répétée B fois de façon à estimer une variance à partir des B courbes de survie estimées par Kaplan-Meier.

Dans l'appel de la procédure , le choix entre l'une ou l'autre solution s'effectue au moyen de mots clefs. Par exemple, on pourra exécuter :

```
proc iclifetest data=... plots=survival(cl) bootstrap(nboot=500);
proc iclifetest data=... plots=survival(cl) bootstrap(nboot=500 seed=123);
proc iclifetest data=... plots=survival(cl) bootstrap;
proc iclifetest data=... plots=survival(cl) impute(nimse=500);
proc iclifetest data=... plots=survival(cl) impute(nimse=500 seed=123);
proc iclifetest data=... plots=survival(cl) impute;
```

Dans ces exemples, le nombre d'échantillons bootstrappés est contrôlé par l'option `nboot=`, celui des échantillons avec imputation par `nimse=`. Pour ces deux options la valeur par défaut est de 1000. Comme toujours, l'option `seed=` gouverne l'initialisation du générateur de nombres au hasard utilisé pour le tirage de ces deux types d'échantillons. Pour plus de précisions quant aux choix qu'il est possible d'opérer sur le graphe de la fonction de survie, je vous invite à regarder l'exemple intitulé *controling the plotting of survival estimates* qui se trouve dans l'aide de la proc `ICLIFETEST`.

6.3.4 Comparaison de courbes de survie sur données censurées