N^0*d'ordre* :15/2015−*M*/*MT*

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIE DE HOUARI BOUMEDIENNE
FACULTÉ DES MATHÉMATIQUES

MEMOIRE
Présenté pour l'obtention du diplôme de MAGISTER

EN : MATHÉMATIQUES
SPÉCIALITÉ : PROBABILITÉ & STATISTIQUE

Par : Soumia KACI

Sujet

Inférence Statistique Dans Des modèles Actuariels De Durée

Soutenue publiquement, le 01/07/2015, devant le jury composé de :

Mm. Z. GUESSOUM	Maître de Conférences/A	à l'USTHB	Présidente.
Mr. K. BOUKHETALA	Professeur	à l'USTHB	Directeur de Mémoire.
Mr. A. TATACHAK	Maître de Conférences/A	à l'USTHB	Examinateur.

Table des matières

Liste des tableaux

Table des figures

Remerciements

Tout d'abord je tiens à remercier Dieu de m'avoir donnée le courage et la santé pour mener à bien ce travail.

Je tiens à exprimer ma profonde reconnaissance à mon directeur de thèse Mr. *Kamal BOUKHETALA* Professeur à l'USTHB pour m'avoir témoigné de sa confiance en me proposant ce sujet et à le remercier pour ses orientations scientifiques et ses précieux conseils qu'il ma accordés pour l'élaboration de ce mémoire.

J'exprime aussi ma profonde gratitude à Mm. *Zohra GUESSOUM* Maître de conférences à l'USTHB, pour avoir accepté de présider le jury de soutenance.

Je remercie également Mr. *Abdelkader TATACHAK* Maître de conférences à l'USTHB, pour avoir accepter d'examiner cette thèse.

Enfin, je remercie chaleureusement toutes les personnes qui m'ont aidé, et qui ont contribué de proche ou de loin à la réalisation de ce travail.

Résumé

L'analyse de durées de vie est utilisée dans des domaines d'application variés et différentes possibilités ont été proposées pour la modélisation de telles données. Nous nous intéressons dans cette thèse à deux types de modélisation différents, Les modèles de survie et les modèles de régression logistique.
Nous proposons des méthodes d'estimation des paramètres et établissons les propriétés asymptotiques des estimateurs obtenus dans chacun de ces modèles.
Nous étudions ces durées dans le contexte des études longitudinales dans le domaine de l'actuariat de l'assurance (durée de survenue d'un sinistre en fonction des paramètres disponible). Nous cherchons alors à estimer la distribution des temps de survie (fonction de survie), à comparer les fonctions de survie de plusieurs groupes ou à analyser la manière dont des variables explicatives modifient les fonctions de survie.

Mots clés : risque actuariel, facteurs de risque, Modèles de durée, inférence Statistique.

Introduction Générale

L'analyse formalisée des données de durée remonte à l'école anglaise d'arithmétique politique , avec notamment les travaux de John GRAUNT (1620-1674) et William PETTY (1623- 1687) à l'occasion des premières études sur la mortalité en Angleterre au 17éme siècle. Les notions d'espérance de vie et d'espérance de vie résiduelle sont alors définies[11].
La recherche de lois sous-jacentes pour ces phénomènes commence au 19éme siècle avec notamment la formule proposée par Benjamin GOMPERTZ en 1825 pour modéliser la probabilité de décéder à l'âge x :

$$h(x) = a \times b^x$$

Ce modèle (qui est en fait une progression géométrique des taux de décès de raison b) sera complété par William MAKEHAM en 1860 :

$$h(x) = c + a \times b^x$$

L'étude des durées de vie restera longtemps un problème étudié par les démographes et les actuaires, jusqu'à l'apparition de la théorie de la "fiabilité" pour les systèmes physiques. Ainsi W. WEIBULL publie en 1951 dans un journal de mécanique un article où il propose la forme suivante pour la fonction de hasard :

$$h(t) = \lambda\alpha t^{\alpha-1}$$

L'article de WEIBULL aborde notamment l'une des particularités importantes des données de durée, la présence de données tronquées ou censurées.
Deux autres dates importantes doivent être citées : l'article de E. KAPLAN et P. MEIER en 1958 [17] dans lequel ils proposent d'utiliser dans le domaine médical un estimateur non paramétrique permettant d'intégrer les données censurées introduit en 1912 par P. BÖHMER, l'estimateur "PL" de la fonction de survie.
En 1972 David COX publie un article posant les bases d'un cas particulier important de modèle à "hasard proportionnel" faisant intervenir des variables explicatives (exogènes) en spécifiant :

$$h(x) = e^{\beta z} h_0(x)$$

avec β un vecteur de paramètres (inconnu) et h_0 la fonction de hasard de base inconnue ; il s'agit donc d'un modèle semi-paramétrique. Ce modèle de référence a donné lieu à de nombreux développements et variantes : introduction d'une évolution temporelle, prise en compte de dépendance entre les variables observées, stratification de l'effet des covariables, etc.
Enfin, pour clore ce bref panorama historique, on peut mentionner deux évolutions récentes des modèles de durées :

– La problématique des tables prospectives et des modèles bi-dimensionnels , dont la référence fondatrice est LEE et CARTER [1992] [18].

– La quantification de la part non mutualisable du risque de mortalité, via les modèles de mortalité stochastique (CAIRNS et al. [2004] [9]).

L'analyse de durée de vie est utilisée dans de nombreux domaines, comme la médecine, la fiabilité industrielle, l'économie ou encore la psychologie, et l'étude de données issues de ces secteurs se développe depuis plusieurs décennies. Il existe de nombreuses manières de modéliser des données de survie et ce travail s'intéresse à deux modélisations diffiérentes ,à travers les **modèles statistiques** et les **modéles de régression** .

Concernant les modèles statistiques proprement dits, trois approches sont possibles : paramétrique, non-paramétrique et semi-paramétrique.

L'**approche paramétrique** stipule l'appartenance de la loi de probabilité *réelle* des observations à une classe particulière de lois, qui dépendent d'un certain nombre (fini) de paramètres.
L'avantage de cette approche est la facilitation attendue de la phase d'estimation des paramètres, ainsi que de l'obtention d'intervalles de confiance et de la construction de tests. L'inconvénient de la méthode paramétrique est l'inadéquation pouvant exister entre le phénomène étudié et le modéle retenu.

L'**approche non-paramétrique** ne nécessite aucune hypothèse quant à la loi de probabilité réelle des observations et c'est là son principal avantage. Il s'agit dès lors d'un problème d'*estimation fonctionnelle*, avec les ambiguités que cela implique par exemple, la fonction de survie, qui est continue, sera estimée par une fonction discontinue.
L'inconvénient d'une telle approche est la nécessité de disposer d'un nombre important d'observations, le problème de l'estimation d'un paramètre fonctionnel étant délicat puisqu'il appartient à un espace de dimension infinie.

L'**approche semi-paramétrique** est une sorte de *compromis* entre les deux approches précédentes. La loi de probabilité réelle des observations est supposée appartenir à une classe de lois pour partie dépendant de paramètres, et pour partie s'écrivant sous forme de fonction(s) non-paramètrique(s). Relativement récente elle est apparue au cours des années soixante dix−, cette approche est très répandue en analyse de la survie, notamment au travers du modèle de régression de Cox (1972)[10].

Les modèles de régression définis par une égalité liant la variable réponse Y aux covariables X

$$Y = \beta' X + \varepsilon,$$

ε étant une variable aléatoire d'erreur.

Dans **le premier chapitre**, nous introduisons et motivons la notion de durée de survie et présentons les outils de modélisation utilisés dans les études d'analyse de durée de vie. Les processus de comptage et les processus empiriques sont un outil essentiel pour établir de nombreux résultats en analyse des données de survie censurées.

Après avoir défini le cadre de l'étude, nous nous intéressons aux différents types de modèles. Nous commençons, dans **le second chapitre**, par présenter les modèles paramétriques sous les trois formes différentes de risque instantanné ,ainsi une généralisation au modèle de Cox en introduisant un vecteur de covariable,ceci permettra de comparer les durées de survie ainsi que les résultats existants pour l'estimation des paramètres du modèle.

Le troisième chapitre est consacré à l'étude des estimateurs non paramétriques à partir des processus poctuels, qui facilite l'obtention d'un certain nombre de résultats via la théorie des martingales. Nous étudions la construction de chacun de ces estimateurs et nous présentons ses propriétés asymptotiques.

Après une revue des modèles paramétrique et non paramétrique, nous présentons dans **le quatrième chapitre** les modèles semi paramétrique, en premier, nous présentons les modèles à hasard proportionnel, en suite nous nous attachons plus particulièrement au modèle semi paramétrique de COX.

Le cinquième chapitre commence par quelques rappels sur les modèles linéaires , en justifiant notamment le passage aux modèles particuliers. Dans la suite de ce chapitre nous nous attachons plus particulièrement au modèle de régression logistique.

Le dernier chapitre concerne un cas pratique d'une compagnie d'assurances automobile. Dans un premier temps, nous proposerons une segmentation concrète de l'échantillon en fonction des variables explicatives disponibles. Nous pourrons évaluer le profil de risque actuel de la compagnie à l'aide du modèle de régression logistique. Cette application aura ainsi pour ambition de montrer la classification des clients de façon cohérente avec la politique de risque de la compagnie.

Chapitre 1

Définitions, Notations et Propriétés

Sommaire

1.1 Introduction

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement précis. L'événement étudié est le passage irréversible entre deux états. L'événement terminal n'est pas forcément la mort : il peut s'agir de l'apparition d'une maladie, d'une guérison (temps entre le diagnostic et la guérison), la panne d'une machine (durée de fonctionnement d'une machine, en fiabilité) ou la survenue d'un sinistre (temps entre deux sinistres, en actuariat).
L'analyse des données (durées) de survie est l'étude du délai de la survenue de cet événement. On cherche alors à estimer la distribution des temps de survie (fonction de survie), à comparer les fonctions de survie de plusieurs groupes ou à analyser la manière dont des variables explicatives modifient les fonctions de survie.

1.2 Concepts de l'inférence statistique

1.2.1 Le modèle statistique

Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène alèatoire.
Une expèrience statistique consiste à recueillir une observation x d'un élément aléatoire X, à valeurs dans un espace $\mathcal{X}$ et dont on ne connait pas exactement la loi de probabilité P. Des considérations de modélisation du phénomène observé amènent à admettre que P appartient à une famille $\mathcal{P}$ de lois de probabilité possibles.

Définition 1.1 *Le **modèle statistique** associé à une expérience est le triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ où*
- *$\mathcal{X}$ est l'espace des observations, ensemble de toutes les observations possibles.*
- *$\mathcal{A}$ est la tribu des évènements observables associée.*
- *$\mathcal{P}$ est une famille de lois de probabilités possibles définie sur $\mathcal{A}$.*

L'intérêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

1.2.2 Le modèle paramétrique ou non paramétrique

Un modèle paramétrique est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre θ inconnu, de dimension d. Alors, la famille de lois de probabilité possibles pour X peut s'écrire $\mathcal{P} = \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\}$.
Le problème principal est alors de faire de l'inférence statistique sur θ : l'estimer, ponctuellement ou par régions de confiance (intervalles si $d = 1$), et effectuer des tests d'hypothèses portant sur θ. On fait alors de la statistique paramétrique.

La statistique non-paramétrique s'intéresse à l'estimation à partir d'un nombre fini d'observations, d'une fonction inconnue $f \in \Phi$, où Φ est un espace fonctionnel assez large.
Ces dernières années, la théorie de l'estimations s'est développée autour des thèmes suivants :

1. Méthodes de constructions d'estimateurs,
2. Propriétés statistiques de ces estimateurs,
3. Optimalité de ces estimateurs
4. Estimation adaptatives.

1.3 Représentation d'une distribution de survie

La loi de probabilité de la durée de survie peut être définie par l'une des fonctions équivalentes suivantes (chacune de ces fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions)[20].

1.3.1 Définitions

Nous donnons ci-dessous les définitions des principaux outils utilisées en analyse de survie. Pour chacun d'eux, nous précisons sa signification statistique d'une part, son interprétation en épidémiologie d'autre part.

Définition 1.2 *La durée de vie d'un individu est une variable aléatoire (v.a.) X positive et continue. Sa* ***fonction de répartition***

$$F(x) = \mathbb{P}(X \leqslant x)$$

est la probabilité que l'évènement se produise entre 0 et x.
Par la suite F sera supposée dérivable.

Définition 1.3 *La* ***fonction de survie*** *est, pour t fixée, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire*

$$\begin{aligned} S(t) &= \mathbb{P}(X > t) \\ &= 1 - F(t) \qquad t \geq 0 \end{aligned}$$

Remarque 1.1 Il est arbitraire de décider que $S(t) = \mathbb{P}(X \geqslant t)$ ou $S(t) = \mathbb{P}(X > t)$. Cela n'a aucune importance quand la loi de X est continue car $\mathbb{P}(X \geqslant t) = \mathbb{P}(X > t)$. Dans le cas où F a des sauts (quand le temps est discret, par exemple, compté en mois ou en semaine), on utilise les notations suivantes :

$$F^-(t) = P(X < t) \quad et \quad F^+(t) = P(X \leqslant t)$$

où F^- est la limite à gauche et F^+ la limite à droite de F (définitions et notations son identique pour S).

Remarque 1.2 La théorie de la survie ayant son origine dans l'observation et le décompte de décès, le vocabulaire est resté marqué pas les termes : "durée de vie", "décès", "exclu vivant"...Cependant, cette théorie s'applique à divers types d'observations : la durée de vie peut ainsi être l'âge d'une apparition d'une maladie, un délai de séroconversion, le temps de sortie du chômage, etc.

Définition 1.4 *La **densité de probabilité** est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$*

$$F(t) = \int_0^t f(u)du$$

Si la fonction de répartition F admet une dérivée au point t alors

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leqslant X < t + \Delta t)}{\Delta t} = F'(t) = -S'(t)$$

Pour t fixé, la densité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t.

Définition 1.5 *La **fonction de hasard** (ou risque instantané) est par définition :*

$$h(t) = \frac{F(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\ln S(t)$$

où f est la densité de probabilité de X.

Notons que cette fonction n'est pas une densité de probabilité. Le risque instantané est la probabilité que l'évènement se produise en t (sachant qu'il ne s'est pas produit auparavant) :

$$\begin{aligned} h(t)\Delta t &= \frac{\mathbb{P}(t < X \leqslant t + \Delta t)}{\mathbb{P}(X > t)} \\ &= \mathbb{P}(X \in]t, t + \Delta t] / X > t). \end{aligned} \tag{1.1}$$

Il en résulte directement que la fonction de hasard détermine entièrement la loi de X et qu'on a la relation suivante :

$$S(t) = \exp\left(-\int_0^t h(s)ds\right) \tag{1.2}$$

Définition 1.6 *La **fonction de hasard cumulé** est donnée par :*

$$H(t) = \int_0^t h(u)du = -\ln S(t). \tag{1.3}$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé :

$$S(t) = \exp(-H(t))$$

1.4 Quantités associées à la distribution de survie

1.4.1 Moyenne et variance de la durée de survie

Le temps moyen de survie $\mathbb{E}(X)$ et la variance de la durée de survie $\mathbb{V}(X)$ se déduisent par intégrale par partie par les quantités suivantes :

$$\mathbb{E}(X) = \int_0^\infty S(t)dt$$

$$\mathbb{V}(X) = 2\int_0^\infty tS(t)dt - (\mathbb{E}(X))^2$$

Ainsi on peut déduire l'espérance et la variance de n'importe laquelle des fonctions F, S, f, h, H.

1.4.2 Quantiles de la durée de survie

La médiane de la durée de survie est le temps t pour le quel la probabilité de survie $S(t)$ égal à 0.5, c'est-à-dire, la valeur t_m qui satisfait $S(t_m) = 0.5$.
Dans le cas où l'estimateur est une fonction en escalier, il se peut qu'il y'ait un intervalle de temps vérifiant $S(t_m) = 0.5$. Il faut alors être prudent dans l'interprétation, notamment si les deux évńements encadrant le temps médian sont éloignés.

Il est possible d'obtenir un intervalle de confiance du temps médian. Soit $[B_i, B_s]$ un intervalle de confiance au niveau α de $S(t_m)$, alors un intervalle de confiance au niveau α du temps médian t_m est

$$[S^{-1}(B_s), S^{-1}(B_i)].$$

La fonction quantile pour la durée de survie est définie par

$$\begin{aligned} q(p) &= \inf(t : F(t) \geqslant p), \qquad 0 < p < 1, \\ &= \inf(t : S(t) \leqslant 1 - p). \end{aligned}$$

Lorsque la fonction de répartition est strictement croissante et continue alors

$$\begin{aligned} q(p) &= F^{-1}(p), \qquad 0 < p < 1, \\ &= S^{-1}(1 - p) \end{aligned}$$

Le quantile $q(p)$ est le temps où une proportion p de la population a disparu.

1.5 Censure et troncature

Une des caractéristiques des données de survie est l'existence d'observations incomplètes.
En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées (i.i.d.) de durées X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié.

1.5.1 Définitions

Définition 1.7 *La **variable de censure** C est définie par la possible non-observation de l'événement.Si l'on observe C, et non X, et que l'on sait que $X > C$ respectivement $X < C$, $C_1 < X < C_2$), on dit qu'il y a **censure à droite** (respectivement **censure à gauche**, **censure par intervalle**).*

Si l'événement se produit, X est "réalisée". S'il ne se produit pas (l'individu étant perdu de vue, ou bien exclu vivant), c'est C qui est " réalisée". X peut être considérée comme la durée séparant un événement initial A d'un événement terminal B, ou comme la durée pendant laquelle un sujet reste dans un état donné (auquel cas A désigne l'entrée dans cet état et B la sortie de cet état -par exemple le chômage). La censure à droite, dont il sera essentiellement question par la suite, est due à la non observation de B, dont on sait seulement qu'il sera postérieur à la dernière date d'observation du sujet.

Par ailleurs, la censure se distingue de la **troncature** : on dit qu'il y a troncature à droite (respectivement à gauche) lorsque la variable d'intérêt X_i (durée de vie d'un individu) n'est pas observable quand elle est supérieure (respectivement inférieure) à un seuil $c > 0$ fixé. Dans le cas de la censure, on sait que la variable X non observée est supérieure ou inférieure à une valeur C qui, elle, a été observée. La troncature, quant à elle, élimine de l'étude une partie des X_i, ce qui a pour conséquence de faire porter l'analyse uniquement sur la loi de X conditionnellement à l'événement $X < c$ (respectivement $X > c$).

Enfin, le mécanisme de censure est habituellement supposé être indépendant de l'événement étudié : on parle de censure **non-informative (ignorable)**. En pratique, cela veut dire que les individus ne doivent pas être censurés parce qu'ils ont un risque de décès particulièrement élevé (ou faible). En d'autres termes, les individus exclus-vivants ou perdus de vue à une date t doivent être représentatifs des individus encore à risque à cet instant t.

Si la censure est **informative**, alors l'expression classique de la vraisemblance ne correspond plus à une vraisemblance complète, mais à une vraisemblance partielle qui peut être utilisée pour des inférences, bien qu'il y ait une perte d'efficacité des estimateurs produits (car toute l'information n'est pas utilisée). Ainsi, la censure informative est à l'origine d'un biais lors de l'analyse standard basée sur la vraisemblance. (Kalbfleisch et Prentice, 1980 ; Schluchter, 1992).

1.5.2 Cractéristiques

Définition 1.8 *La censure est dite **non-aléatoire de type** I si, étant donné un nombre positif fixé c et un n-échantillon $X_1, ..., X_n$, les observations consistent en (T_i, δ_i), où*

$$\begin{cases} T_i = X_i \wedge c \\ \delta_i = 1_{\{X_i \leqslant c\}}. \end{cases}$$

Exemples : test de l'efficacité d'une molécule sur un lot de souris, les souris survivantes étant sacrifiées au bout d'un temps déterminé c ; observation de la durée de fonctionnement de n machines au cours d'une expérience de durée c.

Remarque 1.3 Bien que similaires dans l'écriture de leur définition, la censure non-aléatoire de type I à droite et la troncature à droite doivent être distinguées : en effet, l'inférence statistique diffère grandement, selon qu'elle s'applique à l'un ou l'autre de ces deux types de données de survie. Ainsi, si nous considèrons n observations indépendantes et censurées à droite, la vraisemblance retenue lors de l'étude statistique sera le produit d'un nombre aléatoire (inférieur ou égal à n)de facteurs.
En revanche,si nous considérons maintenant n observations indépendantes et tronquées à droite, nous étudierons une vraisemblance qui sera le produit d'un nombre fixe (exactement n) de facteurs.

Définition 1.9 *La censure est dite **aléatoire de type I** si, étant donné un n-échantillon $X_1,...,X_n$, il existe une v.a. n-dimensionnelle $(C_1,...,C_n)$ de $(\mathbb{R}^+)^n$ telle que les observations consistent en (T_i,δ_i), où*

$$\begin{cases} T_i = X_i \wedge C_i \\ \delta_i = 1_{\{X_i \leqslant C_i\}}. \end{cases}$$

Exemple : lors d'une expérience biologique, on s'intéresse à une cause de décès qui a lieu au bout d'un temps X, et l'on désire étudier la loi de X ; cependant, une autre cause de décès peut intervenir auparavant, et donc empêcher l'observation de X par un mécanisme de censure à droite.

Définition 1.10 *La censure est dite **de type II** si, étant donné un nombre positif fixé r et un n-échantillon $X_1,...,X_n$, les observations consistent en (T_i,δ_i), où*

$$\begin{cases} T_i = X_i \wedge X_{(r)} \\ \delta_i = 1_{\{X_i \leqslant X_{(r)}\}}. \end{cases}$$

où $X_{(1)} < X_{(2)} < ... < X_{(n)}$

Exemples : test de l'efficacité d'une molécule sur un lot de souris, la durée de l'étude correspondant au temps que mettent r souris à mourir ; observation de la durée de fonctionnement de n machines tant que r d'entre elles ne tombent pas en panne.

1.6 La fonction de vraisemblance

Considérons le cas d'une censure aléatoire à droite C indépendante de la durée d'intérêt X (hypothèse d'identifiabilitè).

Supposons que les variables X et C ont pour densités respectives f et g et pour survie S et G. La distribution de X est définie par un paramètre de dimension finie. Toute l'information est contenue dans le couple (T_i,δ_i), où $T_i = \min(X_i,C_i)$ est la durée observée, et l'indicateur de censure $\delta_i = 1_{\{X_i \leqslant C_i\}}$. Ainsi, la contribution à la vraisemblance pour l'individu i est

$$\begin{aligned} L_i &= P(T_i \in [t_i, t_i+\Delta t], \delta_i = 1 \mid \theta)^{\delta_i} \times P(T_i \in [t_i, t_i+\Delta t], \delta_i = 0 \mid \theta)^{1-\delta_i} \\ &= P(X_i \in [t_i, t_i+\Delta t], C_i \geqslant X_i \mid \theta)^{\delta_i} \times P(C_i \in [t_i, t_i+\Delta t], C_i < X_i \mid \theta)^{1-\delta_i} \\ &= [f(t_i \mid \theta) G(t_i^-)]^{\delta_i} \times [g(t_i) S(t_i \mid \theta)]^{1-\delta_i} \end{aligned}$$

Par l'hypothèse (de censure non informative), le paramètre d'intérêt θ n'apparaît pas dans la loi de censure (Il existe des mécanismes indépendants et informatifs). La partie utile de la vraisemblance se réduit alors à

$$L = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i} S(t_i \mid \theta)^{1-\delta_i}$$

Remarque 1.4 Notons que la présence de données censurées doit être prise en compte dans l'écriture de la vraisemblance. En effet, en raisonnant sur le sous échantillon de données non censurées. La vraisemblance est

$$\overline{L} = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i}$$

L'estimateur obtenu en maximisant $\overline{L}$ est asymptotiquement biasé.

Remarque 1.5 Si les observations ne sont pas identiquement distribuées(ex : incorporation de covariables), on peut généraliser l'expression précédante en introduisant un indice pour f et S.

Remarque 1.6 Dans le cas d'une censure non aléatoire, on obtient également la vraisemblance L.

Remarque 1.7 Dans le cas d'une censure aléatoire à droite de type II, la vraisemblance est la suivante :

$$\widetilde{L} = \frac{n!}{k!(n-k)!} \prod_{i=1}^{n} f(t_i \mid \theta) \times S(t_k \mid \theta)^{n-k}$$

1.6.1 Vraisemblance dans un modèle de survie censuré

L'objet de cette section est de déterminer la forme générale de la vraisemblance d'un modèle de durée censuré. En pratique on peut être confronté à une censure droite (si X est la variable d'intérêt, l'observation de la censure C indique que $X \geq C$) ou à une censure à gauche (l'observation de la censure C indique que $X \leq C$) ; les deux types de censure peuvent s'observer de manière concomitante.

Soit X^0 une durée de vie aléatoire. On suppose que la loi P_{X^0} de X^0 appartient à une famille de lois de probabilité $\mathcal{P} = \{P_\theta \in \Theta\}$ où $\Theta \in \mathbb{R}^p$. La vraie loi de X^0 est ainsi notée P_{θ_0} ,où $\theta_0 \in \Theta$.

Notons $f_{X^0;\theta}(.)$, $F_{X^0;\theta}(.)$, $S_{X^0;\theta}(.)$, $h_{X^0;\theta}(.)$, $H_{X^0;\theta}(.)$ les densité, fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de vie X, sous la loi P_X.

La variable de censure C est supposée indépendante de la variable X et sa loi est supposée ne pas dépendre du paramètre θ ; on dit que la loi de la censure C est non informative. On note $f_C(.), F_C(.), S_C(.)$ les densité, fonction de répartition et fonction de

répartition de la variable C.

Les observations sont donc des réalisations de $T = min(X^0;C)$ et de l'indicatrice de censure $\delta = 1_{X^0 \leq C}$. Notons $(T_i, \delta_i)_{i\in\{1,..,n\}}$ un échantillon des variables (T,C). L'estimation de θ_0 à partir des observations peut être effectuffée par la méthode du maximum de vraisemblance.

La vraisemblance associée à l'échantillon $(T_i, \delta_i)_{i\in\{1,..,n\}}$ s'écrit sous la forme

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} (f_{X^0;\theta}(T_i) S_C(T_i))^{\delta_i} (S_{X^0;\theta}(T_i) f_C(T_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} (h_{X^0;\theta}(T_i)^{\delta_i} S_{X^0;\theta}(T_i)) S_C(T_i))^{\delta_i} f_C(T_i)^{1-\delta_i}, \end{aligned}$$

en utilisant les relations liant les densité, fonction de survie et fonction de risque instantané.
Sous l'hypothèse de censure non informative, on remarque qu'il est équivalent de chercher l'estimateur du maximum de vraisemblance de θ en maximisant l'expression

$$\prod_{i=1}^{n} h_{X^0;\theta}(T_i)^{\delta_i} S_{X^0;\theta}(T_i)).$$

1.6.2 Vraisemblance dans un modèle de survie censuré avec covariables

On considère un modèle de prise en compte de covariables à risque multiplicatif. On suppose que la loi conditionnelle $P_{X^0|X}$ de X^0 sachant X appartient à une famille de lois de probabilité $\mathcal{P}_X = \{P_\theta \in \Theta\}$ où $\Theta \in \mathbb{R}^p$. La vraie loi de X^0 sachant X est ainsi notée $P_{\theta_0,X}$, où $\theta_0 \in \Theta$.

Notons les densité, $f_{X^0|X;\theta}(.)$, $F_{X^0|X;\theta}(.)$, $S_{X^0|X;\theta}(.)$, $h_{X^0|X;\theta}(.)$, $H_{X^0|X;\theta}(.)$ fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de vie X^0, sous la loi $P_{\theta,X}$.

On suppose que la loi de X, de densité f^X, ne dépend pas du paramètre θ. De la même façon que dans le cas où les covariables n'interviennent pas, on considère que la loi de X^0 conditionnelle à X est indépendante de la loi de C conditionnelle à X et que la loi de C est non informative pour le paramètre θ.

Avec les mêmes notations que précédemment, la vraisemblance associée à l'échantillon $(T_i, \delta_i, X_i)_{i\in\{1,..,n\}}$ s'écrit, grâce à la formule de Bayes, sous la forme :

$$L_n(\theta) = \prod_{i=1}^{n} \Big(f_{X^0|X;\theta}(T_i) S_{C|X}(T_i) \Big)^{\delta_i} \Big(S_{X^0|X;\theta}(T_i) f_{C|X}(T_i) \Big)^{1-\delta_i} f_X(X_i).$$

Les lois de X et de C conditionnellement à X ne dépendant pas du paramètre θ, l'estimateur du maximum de vraisemblance de θ peut donc être obtenu en maximisant l'expression

$$\prod_{i=1}^{n} h_{X^0|X;\theta}(T_i)^{\delta_i} S_{X^0|X;\theta}(T_i).$$

1.7 Processus de comptage et processus empiriques

Les processus de comptage et la théorie des processus empiriques fournissent des méthodes adaptées pour étudier les données de survie censurées. Cette approche a été développée par Aalen (1975)[1] et le lecteur peut se référer aux ouvrages Fleming Harrington (1991)[14] et Andersen Gill (1982)[5] pour une étude plus complète.

Nous commençons ici par rappeler des notions sur les processus qui seront nécessaires par la suite.

1.7.1 Rappels sur les processus

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé.[7]

Définition 1.11 ***Un processus stochastique réel*** *est une famille de variables aléatoires réelles $X = (X(t))_{t\in\Gamma}$ indexé par un ensemble Γ définies sur le même espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.*
L'ensemble Γ indique en général le temps et vaut habituellement N (processus discrets) ou $\mathbb{R}^+$ (processus continus). On s'intéressera par la suite à des processus continus.

Définition 1.12 *Pour un processus stochastique, les fonctions définies pour $\omega \in \Omega par X(.,\omega)$: $\mathbb{R}^+ \longrightarrow \mathbb{R}$ sont appelées **trajectoires** de X. Un processus sera dit **continu à droite, à variation bornée, croissant, ayant des limites à droite** si l'ensemble des trajectoires ayant la propriété correspondante est de probabilitéé 1.*

Un processus X est donc une application de $\mathbb{R}^+ \times \Omega$ dans $\mathbb{R}$, et $X(t,\omega)$ désigne la valeur de la variable aléatoire $X(t)$ en la réalisation ω.

Définition 1.13 *$\mathcal{B}(\mathbb{R}^+)$ désigne la tribu des boréliens sur $\mathbb{R}^+$.*
*Le processus X est dit **mesurable** si $X : (\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+)) \times (\Omega, \mathcal{F}) \longrightarrow \mathbb{R}$ est une application mesurable.*

Définition 1.14 *Un processus stochastique est dit :*
- ***intégrable** si $\sup_{t\in\mathbb{R}^+} \mathbb{E}|X(t)| < +\infty$,*
- *de **carré intégrable** si $\sup_{t\in\mathbb{R}^+} \mathbb{E}(X(t))^2 < +\infty$,*
- ***borné** s'il existe une constante $M \in \mathbb{R}^+$ telle que*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}^+} |X(t)| < M\right) = 1$$

Les définitions suivantes vont permettre d'établir une formulation rigoureuse du concept d'information s'accroissant avec le temps.

Définition 1.15 *Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, **la filtration** est une famille croissante de sous tribus de $\mathcal{A}$, noté par $(\mathfrak{F}_t, t \geq 0)$. La tribu $\mathfrak{F}_t$ est une description mathématique de toute l'information dont on dispose à l'instant t. Cette information nous permet d'attribuer des probabilités cohérentes aux évènements pouvant intervenir.*

Définition 1.16 *Un processus $\{X_t, t \geq 0\}$ est dit **adapté** à la filtration $(\mathfrak{F}_t, t \geq 0)$ si pour chaque t, X_t est $\mathfrak{F}_t$-mesurable. Un processus adapté est celui pour lequel une description probabiliste est réalisable.*
Les processus de comptage et leurs propriétés seront particulièrement utiles dans la suite
.

Définition 1.17 *Un **processus de comptage** est un processus stochastique $(N_t)_{t \in \mathbb{R}^+}$ adapté à une filtration $(\mathfrak{F}_t)_{t \in \mathbb{R}^+}$ tel que $N(0) = 0, N(t) < +\infty$ p.s. et dont les trajectoires sont avec probabilité 1 continues à droite, constantes par morceaux avec des sauts de taille 1.*

Dans la plupart des applications, comme le terme "processus de comptage" le suggère, $N(t) - N(s)$ représentera le nombre d'événements intervenant dans l'intervalle $]s,t]$. Le processus de Poisson est un des exemples les plus classiques.

Exemple 1.1 Soit X^0 et C deux variables aléatoires positives indépendantes de lois continues. Notons $T = \min(X^0, C)$ l'observation censurée du temps de survie X^0 et $\delta = 1_{\{T \leqslant C\}}$. Alors le processus défini pour $t \geq 0$ par

$$N(t) = 1_{\{T \leq t, \delta = 1\}} = \delta 1_{\{T \leq t\}}$$

est un processus de comptage appelé *processus de comptage des échecs*. De la même façon, le processus défini pour $t \geq 0$ par

$$Y(t) = 1_{\{T \geq t\}}$$

est un processus stochastique appelé *processus à risque*.

Nous définissons maintenant l'intégrale de *Lebesgue-Stieljes* qui nous sera utile lorsque nous intégrerons contre un processus de comptage. Elle se base sur la bijection entre les fonctions croissantes continues à droite et la classe des mesures boréliennes sur $\mathbb{R}$.

Théorème 1.1
Soit $G : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue à droite et posons, pour tout intervalle $[a,b[$ de $\mathbb{R}$ $(a < b), \mu([a,b[) = G(b) - G(a)$. Alors il existe une unique extension de μ à une mesure borélienne sur $\mathbb{R}$.

Soit μ une mesure borélienne sur $\mathbb{R}$ et soit G une fonction, définie sur $\mathbb{R}$ à une constante additive près, par $G(b) - G(a) = \mu(]a,b])$. Alors G est continue à droite et croissante.

Soit $G : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue à droite et posons, pour tout intervalle $[a,b[$ de $\mathbb{R}$ $(a < b), \mu([a,b[) = G(b) - G(a)$. Alors il existe une unique extension de μ à une mesure borélienne sur $\mathbb{R}$.

On a alors les relations suivantes, faciles à établir, entre la fonction croissante continue à droite G et sa mesure borélienne associée μ. La notation $G(t_-) = \lim_{x\to t} G(x)$ est utilisée.

Proposition 1.1
- $\mu(]a,b[) = G(b-) - G(a)$,
- $\mu([a,b]) = G(b) - G(a-)$,
- $\mu([a,b[) = G(b-) - G(a-)$,
  $\mu(\{a\}) = G(a) - G(a-)$,
  G est une fonction continue en a si et seulement si $\mu(\{a\}) = 0$.

La définition d'intégrale de Lebesgue-Stieljes est basée sur la notion plus générale d'intégrale de Lebesgue grâce à la correspondance établie précédemment.

Définition 1.18 *Soit $f : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction borélienne, $G : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue à droite, et μ la mesure borélienne relative à G.*
Pour un ensemble borélien $A \subseteq \mathbb{R}$, on définit ***l'intégrale de Lebesgue-Stieljes*** *$\int_A f dG$ par $\int_A f d\mu$.*

Cette définition assez abstraite permet d'établir des formules plus explicites quand la fonction G a certaines propriétés. Notamment, quand G est une fonction en escalier, elle comporte un nombre au plus dénombrable de sauts notés ici $\{x_1, x_2, ..., \}$ tels que $\delta G(x_n) = G(x_n) - G(x_n-) > 0$. La mesure μ associée sera alors discrète et strictement positive aux points $x_1, x_2, ...$, d'où la formule, pour A un borélien de $\mathbb{R}$

$$\int_A f dG = \sum_{n; x_n \in A} f(x_n)\delta G(x_n).$$

La notation sur un intervalle de l'intégrale $\int_s^t f dG$ pouvant mener à des ambiguïtés si s ou t est un point de discontinuité de G, nous utiliserons la convention

$$\int_s^t f dG = \int_{]s,t]} f dG.$$

Dans cette configuration, on note que l'intégrale de Lebesgue-Stieljes permet d'utiliser une notation simple pour une somme de termes dénombrables.

Si la fonction G a une dérivée g en chaque point de l'intervalle $]s,t]$, alors $\mu(]s,t]) = \int_s^t g(x)dx$ et μ est absolument continue par rapport à la mesure de Lebesgue. On a alors

$$\int_s^t f(x)dG(x) = \int_s^t f(x)g(x)dx.$$

1.7.2 Rappels de théorie des processus empiriques

Nous rappelons ici quelques éléments de théorie des processus empiriques dont nous aurons besoin par la suite. Le lecteur peut se référer à Shorack Wellner (1986)[22], van der Vaart Wellner (1996)[25] et van der Vaart (1998)[26] pour une étude plus complète.
Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans un espace mesurable $(\mathcal{X}, \mathcal{A})$ de même loi de X notée P_X.

Définition 1.19 ***La mesure empirique*** $\mathbb{P}_n$ *associée à* X_n *est la mesure définie sur les boréliens par*

$$\mathbb{P}_n(B) = \frac{1}{n} card\{i \in \{1,...,n\}; X_i \in B\}$$

On la note $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$*, où* δ_a *désigne la mesure de Dirac au point a, c'est la mesure aléatoire qui met un poids* $1/n$ *à chaque observation* X_i*.*

Soit $\mathcal{S}$ un ensemble de fonctions mesurables $f : \mathcal{X} \longrightarrow \mathbb{R}$ et P_X-intégrables, alors la mesure empirique permet de définir une application de $\mathcal{S}$ dans $\mathbb{R}$ donnée par

$$f \longmapsto \mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i).$$

On utilise la notation $Qf := \int f dQ$ pour une fonction mesurable f et une mesure Q. Notons si X est à valeurs réelles pour $f_x = 1_{]-\infty,x]}$, $Qf_x = Q(]-\infty,x])$, et donc on retrouve pour $Q = P_X$ la fonction de répartition de X. On appelle *fonction de répartition empirique* $\mathbb{F}_n$ la fonction de répartition (aléatoire) associée à la mesure empirique, définie par

$$\mathbb{F}_n(x) = \mathbb{P}_n(x) f_x = n^{-1}\sum_{i=1}^n 1_{\{X_i \leq x\}}$$

Définition 1.20 ***Le processus empirique*** $\mathbb{G}_n$ *associé à* (X_n) *et* $\mathcal{S}$ *est l'application*

$$f \longmapsto \mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n(x) - P_X)(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i)) - \mathbb{E}(f(X)).$$

On l'identifiera fréquemment à la mesure $\mathbb{G}_n = n^{-1/2}\sum_{i=1}^n (\delta_{X_i} - P_X)$*.*

Pour une fonction donnée f, sous l'hypothèse d'existence de $P_X f$ et l'hypothèse $P_X(f^2) < +\infty$, la loi forte des grands nombres et le théorème centrale limite donnent les convergences

$$\mathbb{P}_n(f) \xrightarrow{p.s} P_X f \qquad et \qquad \mathbb{G}_n(f) \xrightarrow{\mathcal{F}} \mathcal{N}(0, P_X(f - P_X f)^2).$$

La théorie des processus empiriques s'intéresse au cas plus général de la convergence uniforme de ces processus sur des classes de fonctions. Les théorèmes de Glivenko-Cantelli et Donsker donnent une première extension uniforme sur la classe de fonctions $I = \{1_{]-\infty,x]}, x \in \mathbb{R}\}$ pour les variables aléatoires à valeurs réelles indépendantes.

Théorème 1.2 ***(Glivenko-Cantelli)*** *Soit* (X_n) *une suite de variables aléatoires réelles indépendantes de même fonction de répartition* F_X *et de fonction de répartition empirique* $\mathbb{F}_n$*, alors*

$$\sup_{t\in\mathbb{R}} |\mathbb{F}_n - F| \xrightarrow[n\to+\infty]{p.s} 0.$$

On peut réécrire la conclusion de ce théorème sous la forme $\sup_{f\in I} |\mathbb{P}_n f - P_f| \xrightarrow[n\to+\infty]{p.s} 0$. On note $||Qf||_{\mathcal{S}} = \sup_{f\in\mathcal{S}} |Qf|$. Cela permet d'introduire la définition suivante.

Définition 1.21 *Une classe $\mathcal{S}$ de fonctions $f : \mathcal{X} \longrightarrow \mathbb{R}$ est appelée P_X-**classe de Glivenko-Cantelli** si elle vérifie*

$$||\mathbb{P}_n f - Pf||_{\mathcal{S}} \xrightarrow[n\to+\infty]{p.s} 0$$

La classe I est donc un premier exemple de P_X-classe de Glivenko- Cantelli. Supposons maintenant, afin d'étendre le théorème central limite à une version uniforme, ou fonctionnelle, que

$$\text{pour tout } x \in \mathcal{X}, \qquad \sup_{f\in\mathcal{F}} |f(x) - P_X f| < +\infty.$$

Sous cette condition, le processus empirique $\mathbb{G}_n$ peut être vu comme un élément de l'espace $\ell^\infty(\mathcal{F})$ des fonctions bornées de $\mathcal{F}$ dans $\mathbb{R}$.

Définition 1.22 *Une calsse $\mathcal{S}$ de fonctions $f : \mathcal{X} \longrightarrow \mathbb{R}$ est appelée P_X-**classe de Donsker** si elle vérifie*

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \xrightarrow[n\to+\infty]{l} \mathbb{G} \qquad dans \qquad \ell^\infty(\mathcal{F}),$$

où la limite $\mathbb{G}$ est un processus gaussien tendu centré de fonction de covariance $cov(\mathbb{G}f_1, \mathbb{G}f_2) = P_X f_1 f_2 - P_X f_1 . P_X f_2$.

On remarque que le lemme de Slutsky permet de montrer que toute classe de Donsker est une classe de Glivenko-Cantelli. Réciproquement, toute classe de Glivenko-Cantelli n'est pas une classe de Donsker, mais beaucoup d'exemples peuvent se trouver parmi les classes de Glivenko-Cantelli.
Un premier exemple de classe de Donsker est obtenu grâce au théorème du même nom.

Théorème 1.3 *Soit X_n une suite de variables aléatoires réelles indépendantes de même fonction de répartition F_X et de fonction de répartition empirique $\mathbb{F}_n$, alors la suite des processus empiriques $\sqrt{n}(\mathbb{F}_n - F_X)$ converge en loi dans l'espace des fonctions continues à droite avec limite à gauche sur $\mathbb{R}$ vers un processus gaussien $\mathbb{G}_{F_X}$ tendu centré de fonction de covariance au point (s,t) égale à $F_X(s \wedge t) - F_X(s)F_X(t)$.*

Ainsi, si $\mathcal{X} = \mathbb{R}$, l'ensemble $I = \{1_{]-\infty,x]}, x \in \mathbb{R}\}$ est une P_X-classe de Donsker. van der Vaart Wellner (1996)[25] et van der Vaart (1998)[26] donnent d'autres exemples de classes de Donsker. L'ensemble des fonctions de variation uniformément bornée forme une classe de Donsker.
Soit une classe paramétrique de fonctions $\{f_t, t \in T$, où T est un ensemble borné de $\mathbb{R}^d$. S'il existe m fonction mesurable telle que $m(X)$ admet un moment d'ordre r et pour tout $s,t|f_s(x) - f_t(x)| \leq m(x)||s-t||$, alors cet ensemble est une P_X-classe de Donsker. Les classes de Sobolev sont également des classes de Donsker.

D'autre part, certaines opérations sur les classes de fonctions permettent de préserver la propriété de Donsker. Tout d'abord, si $\mathcal{S}$ est une classe de Donsker, alors tout sous-ensemble de $\mathcal{S}$, l'adhérence et l'enveloppe convexe symétrique de $\mathcal{S}$ sont des classes de Donsker. Si $\mathcal{S}$ et $\mathcal{T}$ sont des classes de Donsker telles que $||P_X||_{\mathcal{S}\cup\mathcal{T}} < \infty$ alors $\mathcal{S} \wedge \mathcal{T} = \{f \wedge g; f \in \mathcal{S}, g \in \mathcal{T}\}, \mathcal{S} \vee \mathcal{T}$ (défini de la même façon),$\mathcal{S} + \mathcal{T}$ et $\mathcal{S} \cup \mathcal{T}$ sont également

des classes de Donsker. Si $\mathcal{S}$ et $\mathcal{T}$ sont des classes de Donsker uniformément bornées alors $\mathcal{S}.\mathcal{T}$ est encore une classes de Donsker. Si $\mathcal{S}$ est Donsker et vérifie $||P_X||_{\mathcal{S}} < \infty$ et l'existence de $\delta > 0$ tel que pour tout $f \in \mathcal{S}$, $f > \delta$ alors $\{1/f, f \in \mathcal{F}\}$ est une classe de Donsker.Pour des théorèmes plus généraux et d'autres exemples nous pourrons se référer à van der Vaart Wellner (1996)[25].

1.7.3 Rappels sur les martingales à temps continu

Certaines méthodes utilisant la théorie des martingales permettent l'étude des propriétés des estimateurs dans le cadre de données de vie censurées.
Nous rappelons donc la définition de martingale à temps continu et donnons une application classique aux données de survie. Le lecteur intéressé par une approche plus complète pourra se référer à Fleming Harrington (1991)[14].

Soit $X = (X(t))_{t\geq 0}$ un processus stochastique continu à droite avec limites à gauche et $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ une filtration.

Définition 1.23 *X est appelé **martingale** adaptée à la filtration $(\mathcal{F}_t)_{t\in\mathbb{R}}$ ou $\mathcal{F}_t$-martingale si*

- *X est adapté à $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$,*
- *pour tout $t \in \mathbb{R}^+, \mathbb{E}|X(t)| < +\infty$,*
- *pour tout $s,t \in \mathbb{R}^+, \mathbb{E}(X(t+s) \mid \mathcal{F}_t) = X(t)$ p.s.*

*Si la dernière condition est remplacée par $\mathbb{E}(X(t+s) \mid \mathcal{F}_t) \geq X(t)$ p.s., X est appelé **sous-martingale**.*

*Si la dernière condition est remplacée par $\mathbb{E}(X(t+s) \mid \mathcal{F}_t) \leq X(t)$ p.s., X est appelé **sur-martingale**.*

Proposition 1.2 Soit X une martingale adaptée à $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$. Alors $\mathbb{E}(X(t) \mid \mathcal{F}_{t^-}) = X(t^-)$ p.s.

Exemple 1.2 Reprenons l'exemple 1.1. Introduisons la filtration définie par

$$\mathcal{F}_t = \sigma\{N(s), (1-\delta)1_{X^0\leq s}; 0s \leq t\}.$$

Notons h la fonction de risque instantanée de X^0 et définissons le processus M sur $\mathbb{R}^+$ par

$$M(t) = \delta 1_{X^0\leq s} - \int_0^t 1_{X^0\geq u} h(u)du.$$

Alors M est une martingale adaptée à la filtration $\mathcal{F}_t$.
Ce résultat est également valable sous des hypothèses plus faibles que l'indépendance de X^0 et la censure C mais il sera suffisant pour les configurations auxquelles nous nous intéressons.

Chapitre 2

Modèles paramétriques

Sommaire

2.1 Introduction

Un modèle paramétrique peut être formulé en précisant la forme de l'une ou l'autre des cinq fonctions équivalentes qui définissent la loi de la durée :h, H, f, S ou F. Souvent,on suppose que la distributions des durées de survie appartient à une famille de loi paramétrique donnée. Néanmoins, on spécifie souvent la forme du risque instantané λ : constant, monotone croissant ou décroissant et en forme de $\cap$ ou de $\cup$.

Les estimateurs des paramètres du modèle sont ensuite obtenus en maximisant la vraisemblance des observations (par l'intermédiaire de méthodes itératives, par exemple l'algorithme de Newton-Raphson).

2.2 Risque instantané constant(loi exponentielle)

La loi exponentielle $\mathcal{E}(\theta)$, qui ne dépend que d'un paramètre θ, est la seule qui admet un risque instantané constant. Cette loi est aussi dite "sans mémoire" car la probabilité de décès pour un individu dans un certain laps de temps est la même quelle que soit sa durée de vie $(i.e : P(X > s+t \mid X > t) = P(X > s))$. Les quantités associées à cette loi sont :

$$\begin{aligned} f(t \mid \theta) &= \theta e^{-\theta t}, \qquad t \geqslant 0 \quad et \quad \theta > 0, \\ h(t \mid \theta) &= \theta, \\ S(t \mid \theta) &= e^{-\theta t}. \end{aligned}$$

Dans certaines applications, on peut découper le temps en plusieurs intervalles et considérer un θ_i différent pour chacun des intervalles (risque est constant sur chaque période mais varie d'une période à une autre).

2.3 Risque instantané monotone

2.3.1 Loi de Weibull

Ce sont des lois qui généralisent la loi exponentielle, et pour les quelles le risque instantané est une puissance de temps. Considérons une loi de Weibull $\mathcal{W}(\alpha, \theta)$; alors

$$\begin{aligned} f(t \mid \alpha, \theta) &= \alpha\theta^{\alpha} t^{\alpha-1} \exp\{-(\theta t)^{\alpha}\}, \qquad t \geqslant 0 \quad et \quad (\theta, \alpha > 0), \\ h(t \mid \alpha, \theta) &= \alpha\theta^{\alpha} t^{\alpha-1}, \\ S(t \mid \alpha, \theta) &= \exp\{-(\theta t)^{\alpha}\} \\ \mathbf{E}(T \mid \alpha, \theta) &= \frac{1}{\theta}\Gamma(1+\frac{1}{\alpha}), \\ \mathbf{V}(T \mid \alpha, \theta) &= (\frac{1}{\theta})^2(\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+\frac{1}{\alpha})). \end{aligned}$$

θ est un parmètre d'échelle et α un parmètre de forme.
Lorsque $\alpha = 1$, on retrouve la loi exponentielle $W(1, \theta) = \mathcal{E}(\theta)$.
Lorsque $\alpha = 2$ et $\theta = 1/2$ ce modèle porte le nom de "modèle de RAYLEIGH" ; il est

utilisé en physique pour modéliser la durée de vie de certaines particules ou le bruit de sortie de certains récepteurs de transmissions.
Si $0 < \alpha < 1$ (Figure2.1), le risque instantané est decroissant de ∞ à 0.
Si $\alpha > 1$, le risque instantané est croissant de 0 à ∞ (Figure2.2).

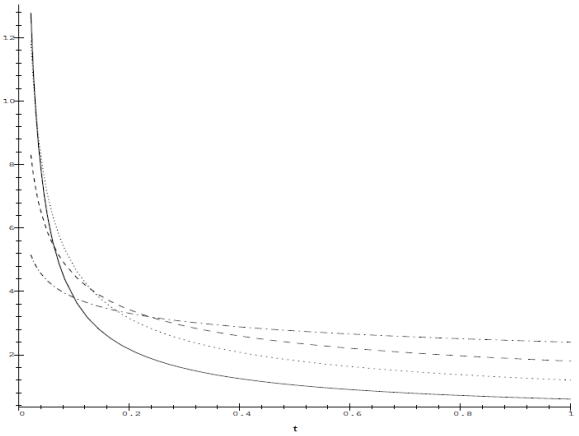

FIGURE 2.1 – Loi de Weibull

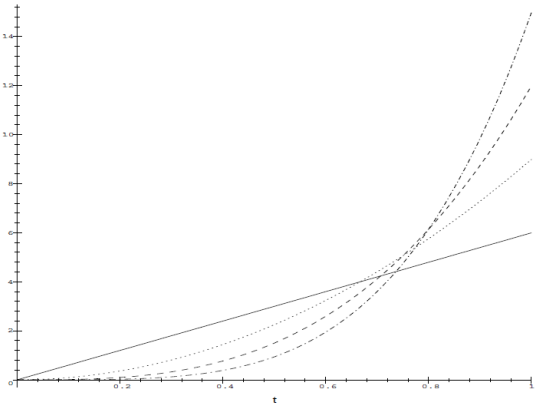

FIGURE 2.2 – Loi de Weibull

2.3.2 Loi Gamma

Le modèle Gamma est une autre généralisation naturelle du modèle exponentiel. Considérons une loi Gamma $G(\alpha,\theta)$; alors

$$f(t \mid \alpha,\theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\theta t}, \qquad t \geqslant 0 \quad et \quad (\theta,\alpha > 0),$$
$$F(t \mid \alpha,\theta) = \frac{1}{\Gamma(\alpha)} \int_0^{\theta t} u^{\alpha-1} e^{-u} du,$$
$$h(t \mid \alpha,\theta) = \frac{f(t \mid \alpha,\theta)}{1 - F(t \mid \alpha,\theta)}$$

avec $\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} e^{-u} du$.Pour $\alpha = 1$, on retrouve la loi exponentielle $\mathcal{E}(\theta)$.
Si $0 < \alpha < 1$, le risque instantané est décroissant de ∞ à $\frac{1}{\theta}$ (Figure2.4).
Si $\alpha > 1$ le risque instantané est croissant de 0 à θ (Figure2.3).

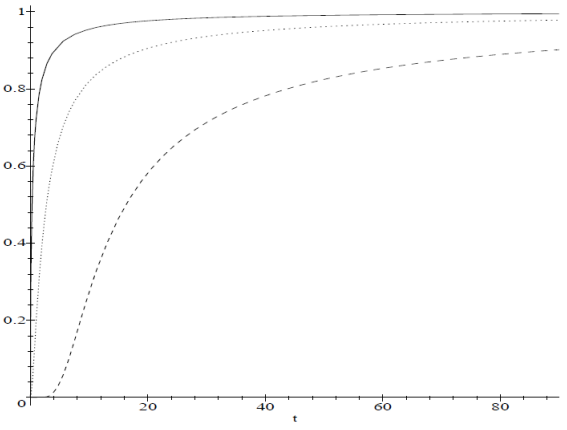

FIGURE 2.3 – Loi Gamma

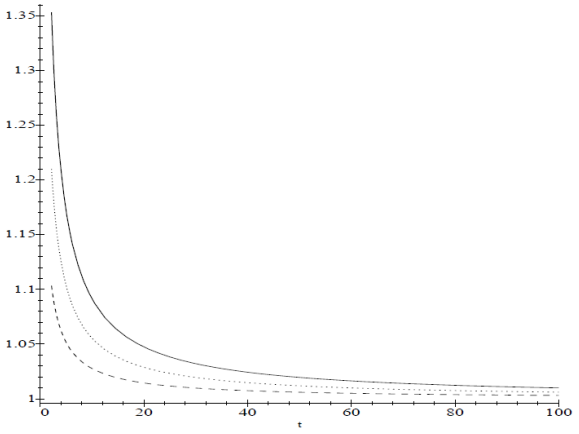

FIGURE 2.4 – Loi Gamma

Les tests d'adéquation ne permettent de distinguer la loi de Weibull de Gamma que lorsque la taille de l'échantillon est trés grande.

2.3.3 Loi de Gompertz Makeham

Il s'agit du modèle de référence pour la construction de tables de mortalité et, dans une moindre mesure, de tables de maintien en arrêt de travail. Une loi de Gompertz Makeham $GM(\alpha,\beta,\gamma)$ est définie par la densité suivante :

$$f(t) = (\alpha+\beta+\gamma)\exp\{-\alpha t - \frac{\beta}{2}(e^{-\gamma t}-1)\};$$

et par la fonction de hasard suivante :

$$h(t) = \alpha + \beta \times \gamma^t$$

En démographie, la forme de cette fonction s'interprète de la manière suivante : le paramètre α représente un taux de décès accidentel (indépendant de l'âge), le terme en $\beta \times \gamma^t$ modélise quant à lui un vieillissement exponentiel (si $\gamma > 1$). Incidemment on retrouve le modèle exponentiel si $\beta = 0$. Par rapport à d'autres modèles, la fonction de Makeham a donc une ambition "explicative", ou "physique", en intégrant explicitement deux causes de décès clairement identifiées.

De manière plus précise, si on considère que le décès peut survenir de deux causes "concurrentes", l'accident et le vieillissement, la date de décès est de la forme $T = T_A \wedge T_V$, T_A (resp. T_V) représentant le décès accidentel (resp. dû au vieillissement). On suppose le décès accidentel modélisé par une loi exponentielle de paramètre a, et le décès associé au vieillissement modélisé par la fonction de hasard de Gompertz $\beta \times \gamma^t$; alors T suit une loi de Makeham. Cela découle immédiatement du fait que la fonction de survie de T est le produit des fonctions de survies de T_A et T_V , et donc les fonctions de hasard s'ajoutent.

Un calcul direct conduit aisément à l'expression de la fonction de survie :

$$S(t) = \exp\left\{-\alpha t - \frac{\beta}{\ln(\gamma)}(\gamma^t - 1)\right\}.$$

Le calcul de l'espérance de T est par contre complexe :

$$\mathbf{E}(t) = \int_0^{+\infty} e^{\{-\alpha t - \frac{\beta}{\ln(\gamma)}(\gamma^t - 1)\}} dt.$$

Mais :

$$S(t) = e^{\frac{\beta}{\ln(\gamma)}} \times e^{-\alpha t} \times e^{-\frac{\beta}{\ln\gamma}\gamma^t}$$

On effectue alors le changement de variable :

$$u = \frac{\beta}{\ln(\gamma)}\gamma^t = \frac{\beta}{\ln(\gamma)} e^{t \times \ln(\gamma)}, \frac{du}{\ln(\gamma)u} = dt$$

qui implique $\left(\frac{\ln(\gamma)}{\beta}u\right)^{1/\ln(\gamma)} = e^t$ puis :

$$\begin{aligned}\mathbf{E}(T) &= e^{\frac{\beta}{\ln(\gamma)}} \times \int_{\frac{\beta}{\ln(\gamma)}}^{+\infty} \left(\frac{\ln(\gamma)}{\beta}u\right)^{-\alpha/\ln(\gamma)} \times e^{-u} \frac{du}{\ln(\gamma) \times u} \\ &= \frac{1}{\ln(\gamma)} \left(\frac{\ln(\gamma)}{\beta}\right)^{-\alpha/\ln(\gamma)} e^{\frac{\beta}{\ln(\gamma)}} \times \int_{\frac{\beta}{\ln(\gamma)}}^{+\infty} u^{-(1+\alpha/\ln(\gamma))} e^{-u} du\end{aligned}$$

Avec le changement de variable $v = u - \frac{\beta}{\frac{\beta}{\ln(\gamma)}}$ on trouve

$$\mathbf{E}(T) = \frac{1}{\beta} \times \int_0^{+\infty} \left(\frac{\ln(\gamma)}{\beta}v + 1\right)^{-(1+\alpha/\ln(\gamma))} \times e^{-v} dv$$

2.3.4 Mélange de deux distributions exponentielle

2.3.4.1 Agrégation de lois

Il arrive souvent en pratique que les durées que l'on observe résultent de l'agrégation de sous populations ayant chacune un comportement spécifique, souvent inobservable. On parle alors d'hétérogénéité.
On suppose ici que la fonction de survie dépend d'un paramètre aléatoire ν , ce paramètre étant distribué selon une loi π . D'un point de vue heuristique, on se trouve en présence de sous-populations à l'intérieur desquelles la loi de survie est homogène et dècrite par la loi de survie conditionnelle au fait que la valeur du paramètre soit ν , $S(t,\nu)$, la loi π décrivant le poids respectif de chaque sous-population dans la population totale.
On a donc la forme suivante pour la fonction de survie initiale de la population totale $S(t) = \int S(t,\nu)\pi(d\nu)$:

$$S(t) = P(T > t) = \mathbf{E}_\nu[P(T > t \mid \nu)] = \int S(t,\nu)\pi(d\nu).$$

La distribution d'hétérogénéité dépend à priori de t, puisque les individus des différentes sous-populations ne sortent pas du groupe à la même vitesse. A la date t, et en supposant la taille de la population infinie, on a ainsi :

$$\pi(d\nu) = \frac{S(t,\nu)}{S(t)}\pi(d\nu)$$

La fonction de hasard à la date t s'écrit alors $h(t) = \int h(t,\nu)\pi_t(d\nu)$. En effet, il suffit de remarquer que

$$\frac{P(T \leqslant t+u \mid T > t)}{u} = \int \frac{P(T \leqslant t+u \mid T > t, \nu)}{u} \pi_t(d\nu),$$

puis de faire tendre u vers 0.Dans le cas particulier où $S(t,\nu) = \exp(-\lambda(\nu)t)$, c'est à dire où chaque sous-population est décrite par une loi exponentielle de paramètre $h(t,\nu) = \lambda(\nu)$, la fonction de survie agrégée s'écrit :

$$S(t) = \int_0^{+\infty} \exp(-\lambda(\nu)t)\pi(d\nu)$$

D'après l'expression ci-dessus de la fonction de hasard s'écrit donc $h(t) = \int \lambda(t)\pi_t(d\nu)$et on en déduit que :

$$\frac{dh(t)}{dt} = -\int \lambda^2(\nu)\pi_t(d\nu) + \left(\int \lambda(t)\pi_t(d\nu)\right)^2$$

En effet, de l'expression de $\pi_t(d\nu) = \dfrac{S(t,\nu)}{S(t)}\pi(d\nu)$ il découle :

$$\frac{\partial}{\partial t}\pi_t(d\nu) = \frac{\frac{\partial}{\partial t}S(t,\nu) \times S(t) - S(t,\nu) \times \frac{d}{dt}S(t)}{S(t)^2}\pi(d\nu)$$

avec $\dfrac{\partial}{\partial t}S(t,\nu) = -\lambda(\nu)S(t,\nu)$ et $\dfrac{S'(t)}{S(t)} = -h(t) = -\int h(\nu)\pi_t(d\nu)$. On en déduit :

$$\frac{\partial}{\partial t}\pi_t(d\nu) = \frac{-\lambda(\nu) \times S(t,\nu)}{S(t)}\pi(d\nu) + \frac{S(t,\nu) \times h(t)}{S(t)}\pi(d\nu) = -\lambda(\nu)\pi_t(d\nu) + h(t)\pi_t(d\nu)$$

En écrivant $\dfrac{d}{dt}h(t) = \int \lambda(\nu)\dfrac{\partial}{\partial t}\pi_t(d\nu)$ on trouve donc finalement :

$$\frac{dh(t)}{dt} = -\int \lambda^2(\nu)\pi_t(d\nu) + h(t)^2$$

Ce qui est le résultat attendu. L'agrégation de fonctions de hasard constantes conduit donc à une fonction de hasard globale décroissante. Ce phénomène s'explique par le fait que les individus ayant une valeur élevée de $\lambda(\nu)$ sortent en premier et il reste donc proportionnellement plus d'individus à $\lambda(\nu)$ faible lorsque le temps s'écoule. Le taux de sortie est donc logiquement décroissant. Ce phénomène porte le nom de "biais d'hétérogénéité", ou "mobile-stable".

2.3.4.2 Exemple :*Mélange de deux lois exponentielles*

La durée est ici une variable exponentielle de paramètre θ_1 avec la probabilité p_1 et θ_2 avec la probabilité $p_2 = 1 - p_1$, soit

$$\begin{aligned}
&S(t) = p_1 e^{-\theta_1 t} + p_2 e^{-\theta_2 t} \qquad (0 < p_1 < 1, \theta_1 > \theta_2 > 0);\\
&f(t) = p_1\theta_1 e^{-\theta_1 t} + p_2\theta_2 e^{-\theta_2 t}\\
&h(t) = f(t)/S(t)
\end{aligned}$$

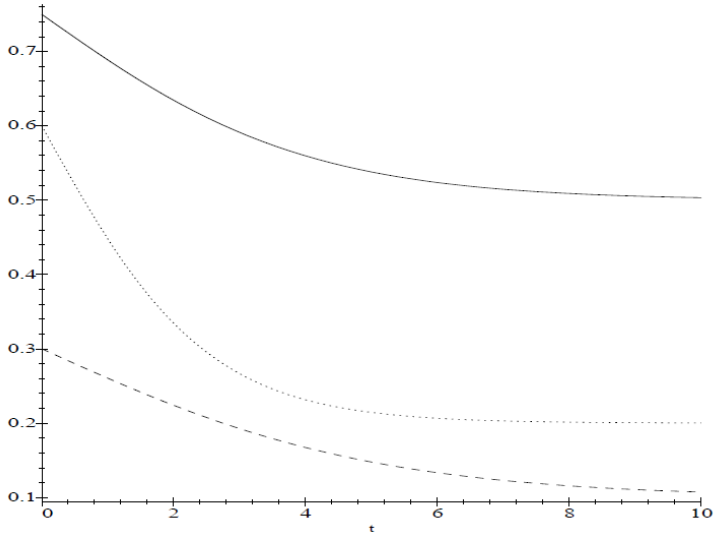

FIGURE 2.5 – Mélange de deux lois exponentielles

Le risque instantanné est décroissant de $c_2 = \frac{p_1}{\theta_1} + \frac{p_2}{\theta_2}$ à $c_1 \frac{1}{\theta_2}$ (Figure 2.5).

2.4 Risque instantané en $\cap$ et $\cup$

2.4.1 Loi de Weibull généralisée

La loi de Weibull est intéressante pour modéliser des risques monotones. Cependant elle devient mal adaptée quand les risques sont en forme de cloche. Une alternative est l'utilisation de la loi de Weibull généralisée $GW(\alpha, \theta, \gamma)$

$$h(t \mid \theta, \nu, \gamma) = \left(1 + (\frac{t}{\theta})^{\nu}\right)^{\frac{1}{\gamma}-1} \frac{\nu}{\gamma\theta^{\nu}} t^{\nu-1} \qquad t \geqslant 0, (\theta, \nu, \gamma > 0),$$

$$S(t \mid \theta, \nu, \gamma) = \exp\left[1 - \left(1 + (\frac{t}{\theta})^{\nu}\right)^{\frac{1}{\gamma}}\right].$$

Pour $\gamma = 1$, on retrouve la loi de Weibull $\mathcal{W}(\theta, \nu)$; pour $\gamma = 1$ et $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}(\frac{1}{\theta})$. En faisant varier les paramètres on peut obtenir des risques constants, monotones croissant ou décroissant, avec les formes en $\cap$ et des formes en $\cup$.
Les risques en forme de cloche sont souvent présents dans le domaine du vivant. Par exemple, les risques instantanés en forme de $\cup$ comportent 3 phases : la période de mortalité infantile, la période de risque faible et la période de vieillissement durant lequel le risque augmente.

2.4.2 Loi log normale $LN(\mu,\sigma)$

$$S(t \mid \mu,\sigma) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right) \qquad (\mu \in \Re; \sigma > 0); t \geqslant 0;$$
$$f(t \mid \mu,\sigma) = \frac{1}{\sigma t}\varphi\left(\frac{\ln t - \mu}{\sigma}\right);$$
$$h(t \mid \mu,\sigma) = \frac{f(t \mid \mu,\sigma)}{S(t \mid \mu,\sigma)};$$
$$\mathbf{E}(T) = e^{\mu+\sigma^2/2}, \qquad \mathbb{V}(T) = e^{2\mu+\sigma^2/2}(e^{\sigma 2} - 1).$$

Ici Φ est la fonction de répartition de la loi normale standard,

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} = \Phi'(t).$$

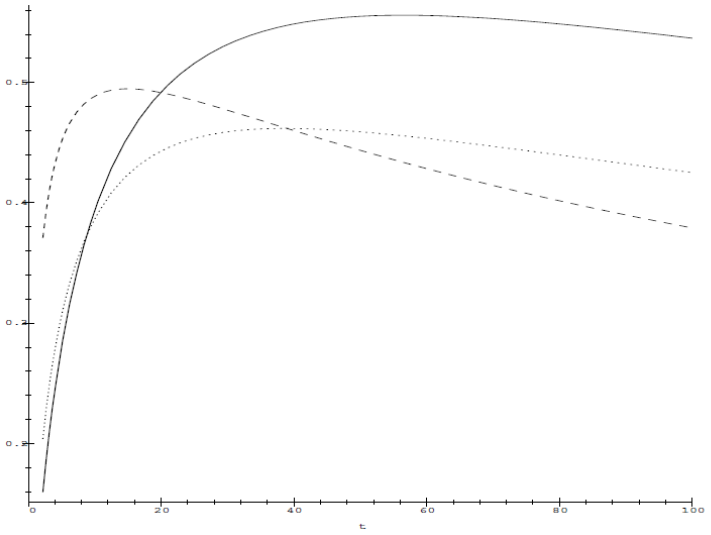

FIGURE 2.6 – Loi log normale

Le risque instantané croît de 0 à sa valeur maximum puis décroit vers 0 ,i.e., il est en forme de $\cap$ (Figure 2.6).

2.4.3 Loi log logistique $LL(\theta,\nu)$

$$S(t \mid \theta,\nu) = \frac{1}{1 + (\frac{t}{\theta})^\nu} \qquad (\theta,\nu > 0),$$
$$h(t \mid \theta,\nu) = \frac{\nu}{\theta^\nu} t^{\nu-1} \left(1 + (\frac{t}{\theta})^\nu\right)^{-1},$$
$$f(t,\theta,\nu) = \frac{\nu}{\theta^\nu} t^{\nu-1} \left(1 + (\frac{t}{\theta})^\nu\right)^{-2}.$$

Pour $0 < \nu \leqslant 1$ la moyenne n'existe pas. Pour $\nu > 1$

$$\mathbf{E}(T) = \theta\Gamma(1+1/\nu)\Gamma(1-1/\nu).$$

La variance existe pour $\nu > 2$

$$\mathbf{Var}(T) = \theta^2\{\Gamma(1+2/\nu)\Gamma(1-2/\nu) - \Gamma^2(1+1/\nu)\Gamma^2(1-1/\nu)\}.$$

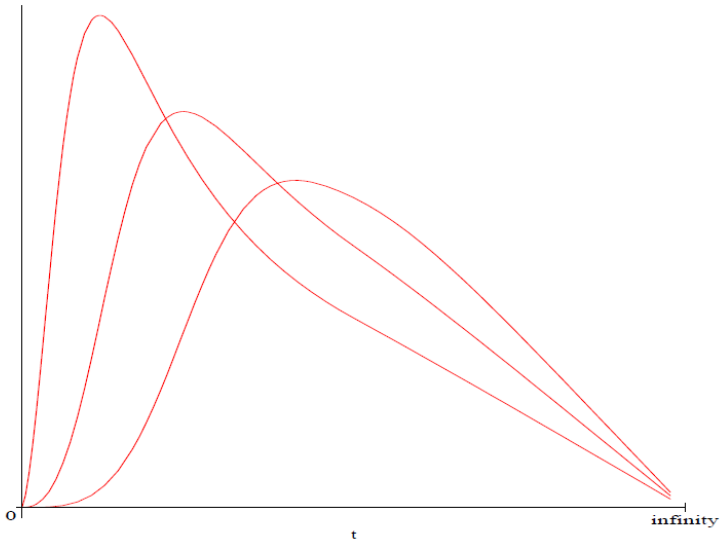

FIGURE 2.7 – Loi log-logistique

Pour $\nu > 1$ le risque instantané croît de 0 à sa valeur maximum puis décroit vers 0 ,c'est à dire qu'il est en forme de $\cap$ (Figure 2.7).

2.5 Introduction de cavariables

Dans l'approche paramétrique, les fonctions d'intérêts peuvent dépendre de covariables explicatives succeptibles d'influencer la survie. En plus d'ajuster les fonctions de survie à différents facteurs, ceci permettra de comparer les durées de survie (l'hypothèse nulle sera l'égalité des distributions de survie). Considérons Z un vecteur de covariables. Notons que ces covariables peuvent dépendre du temps, cependant il est nécessaire de supposer que la valeur des covariables ne change pas entre deux mesures. Afin de simplifier les écritures on supposera dans ce qui suit que les covariables sont fixées au cours du temps. On suppose que les covariables vont modifier les fonctions de risque en suivant un modèle à risques proportionnels "de Cox" (d'autres modèles à risques proportionnels sont possibles), c'est-à-dire

$$h(t \mid Z) = h_0(t)\exp(\beta' Z)$$

où β est le vecteur des coefficients de régression. Les fonctions de survie et de densité correspondant à ces fonctions de risque sont données par

$$S(t \mid Z) = \exp\left(-\int_0^t h(u \mid Z)du\right) = \exp\left(-\int_0^t h_0(u)\exp(\beta' Z)du\right) = S_0(t)^{exp(\beta' Z)}$$

$$f(t \mid Z) = -S'(t \mid Z) = h(t \mid Z)\exp\left(-\int_0^t h(u \mid Z)du\right) = h_0(u)\exp(\beta' Z)du \times S_0(t)^{exp(\beta' Z)},$$

avec $S_0(t) = \exp\left(-\int_0^t h_0(u)du\right)$.
Les paramètres du modèle s'obtiennent simplement par la méthode du maximum de vraisemblance.

2.5.1 Comparaison de deux groupes

Considérons la situation où l'on souhaite comparer les durées de survie de deux groupes A et B. On introduit la covariable suivante,

- $Z = 0$ si l'individu appartient au groupe $A \Longrightarrow \lambda_A(t) = \lambda_0(t)$
- $Z = 1$ si l'individu appartient au groupe $B \Longrightarrow \lambda_B(t) = \lambda_0(t)\exp(\beta)$

Pour comparer les deux groupes, on estime le coefficient de régression β et on teste l'hypothèse nulle $H_0 : \beta = 0$ c'est-à-dire $H_0 : \lambda_A = \lambda_B$: On peut,à cet effet, utiliser les tests du rapport de vraisemblance, de Wald ou du score qui suivent asymptotiquement une loi de $\chi^2(1)$; sous H_0.

Exemple 2.1 Considérons un risque de base suivant une loi de Weibull $\mathcal{W}(\theta,\nu)$; alors

$$
\begin{aligned}
h_0(t) &= \nu\left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1}, \quad t \geq 0 \quad et \quad \theta,\nu > 0,\\
S_0(t) &= exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right),\\
f_0(t) &= \nu\left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1}\exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right).
\end{aligned}
$$

D'après les résultats du début de la section, les fonctions de risque, de survie et de densité dans le cas où il y a des covariables sont

$$
\begin{aligned}
h(t \mid Z) &= nu\left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \times \exp(\beta' Z), \quad t \geq 0 \quad et \quad \theta,\nu > 0,\\
S(t \mid Z) &= exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)^{\exp(\beta' Z)},\\
f(t \mid Z) &= \nu\left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \times \exp(\beta' Z) \times \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)^{\exp(\beta' Z)}
\end{aligned}
$$

Pour $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}(\frac{1}{\theta})$: Ainsi, dans le cas d'un risque suivant une loi exponentielle avec des covariables, on obtient

$$
\begin{aligned}
h(t \mid Z) &= \frac{1}{\theta} \times \exp(\beta' Z), \quad \theta > 0,\\
S(t \mid Z) &= exp\left(-\frac{t}{\theta}\right)^{\exp(\beta' Z)},\\
f(t \mid Z) &= \frac{1}{\theta}\exp(\beta' Z) \times \exp\left(-\frac{t}{\theta}\right)^{\exp(\beta' Z)}
\end{aligned}
$$

2.5.2 Modèles de vie accélerée (Accelerated Failure Time model)

Parmi les modèles de régression, les modèles de vie accélerée sont souvent considérés notamment en fiabilité. Ces modèles peuvent être définis de deux manières. La première représentation des modèles de vie accélérée est donnée par la fonction de survie accélérée :

$$S(t \mid Z) = S_0(te^{\beta' Z})$$

où Z est un vecteur de covariable, β le vecteur des coefficients de régression. Le terme $e^{\beta' Z}$ est un facteur d'accélération car un changement dans les covariables change l'échelle de temps. On peut obtenir une expression de la fonction de risque,

$$h(t \mid Z) = [-\ln(S(t \mid Z))]' = -\frac{[S(t \mid Z)]}{S(t \mid Z)} = -\frac{-e^{\beta' Z} \times h_0(te^{\beta' Z}) \times S_0(te^{\beta' Z})}{S_0(te^{\beta' Z})} = e^{\beta' Z} h_0(te^{\beta' Z})$$

En effet, on a les égalités suivantes,

$$S(t \mid Z) = S_0(te^{\beta' Z}) = \exp(-H(te^{\beta' Z})) = \exp\left[-\int_0^t h_0(ue^{\beta' Z})du\right]$$

Si on suppose que $S_0(t)$ est la fonction de survie de la variable $\exp(\mu+\varepsilon)$ alors $S_0(t) = P(e^{\mu+\varepsilon} > t)$: Ainsi, on obtient que

$$S(t \mid Z) = S_0(te^{\beta' Z}) = P(e^{\mu+\varepsilon} > te^{\beta' Z}) = P(e^{\mu-\beta' Z+\varepsilon} > t) = P(X > t),$$

est la fonction de survie de la variable X où $\log(X) = \mu - \beta' Z + \varepsilon$. En considérant le changement de variable $-\alpha = \beta$, on obtient la deuxième représentation par un modèle de régression log-linéaire pour la durée de survie

$$\log(X) = \mu + \alpha' Z + \varepsilon$$

où X est la durée de survie (pas toujours observée car $T = \min(X, C)$) etε est une variable aléatoire (dans le cas de plusieurs observations, les ε_i sont $i.i.d.$).
Plusieurs lois sont possibles pour les variables ε_i, par exemple,

- $\varepsilon \sim$ loi aux valeurs extrêmes $(f_\varepsilon(y) = \exp(y - e^y))$
- $\varepsilon \sim$ log-logistique
- $\varepsilon \sim$ log-normale
- $\varepsilon \sim$ gamma généralisée

On peut déduire la loi de X et les estimations des paramètres sont obtenues par maximisation de la vraisemblance.

Remarque 2.1 *On peut remarquer que dans le cas des modèles de vie accélerée, pour une covariable $Z > 0$; un coefficient de régression α négatif entraîne un temps de survie plus petit est donc une survie plus faible. Alors que dans le modèle semi-paramétrique de Cox un coefficient de régression α négatif entraîne un risque d'événement plus faible et donc une survie plus grande.*

Chapitre 3

Modèles non paramétriques

Sommaire

3.1 Introduction

On peut souhaiter, dans un certain nombre de situations, ne pas faire d'hypothèse à priori sur la forme de la loi de survie ; on cherche donc à estimer directement cette fonction, dans un espace de dimension infinie ; ce cadre d'estimation fonctionnelle est le domaine de l'estimation non paramétrique.

Sous réserve de disposer de données en quantités suffisantes, on peut alors obtenir des estimations fiables de la fonction de survie, et des fonctionnelles associées.

Dans le contexte usuel d'un échantillon i.i.d. non censurés $(T_1,..,T_n)$, on dispose de l'estimateur empirique de la fonction de répartition $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i \leq t\}}$. Cet estimateur possède un certain nombre de *bonnes propriétés* bien connues : il est sans biais, convergent et asymptotiquement gaussien. Plus précisément, la convergence est uniforme au sens presque sur, et on a le *théorème central limite* suivant :

$$\sqrt{n}(F_n - F) \longrightarrow W$$

Où W est un processus gaussien centré de covariance $\rho(s,t) = F(s) \wedge F(t) - F(s)F(t)$. Ce résultat découle directement du théorème de Donsker dans le cas de la loi uniforme [1] et du fait que $F(T)$ suit une loi uniforme sur $[0,1]$.

L'objectif de l'estimation empirique dans les modèles de durée est de rechercher un estimateur vérifiant des propriétés équivalentes en présence de censure. Pour ce faire, on commence par introduire la présentation des modèles de durée à partir de processus ponctuels, qui facilite ensuite l'obtention d'un certain nombre de résultats via les résultats limite sur les martingales.

3.2 Modèles de durée et processus ponctuels

L'étude d'une durée de survie s'effectue en général en étudiant la loi de la variable X, associée à la fonction de survie S. On considère le processus ponctuel naturellement associé à X, $N(t)$, égal à 0 tant que l'évènement n'a pas eu lieu, puis 1 après :$N(t) = 1_{\{X \leqslant t\}}$. Lorsque l'on prend en compte la censure, on construit de même $N^1(t) = 1_{\{T \leqslant t, D=1\}}$ le processus des sorties non censurées.
Cette approche fait largement appel à la théorie des martingales, dont les résultats essentiels sont cités dans la section..., et quelques propriétés sont rappelés ci-après.

1. Le processus limite étant alors le pont brownien, processus gaussien centré de covariance $s \wedge t - st$

M_t est une $\mathfrak{F}_t$-sur-martingale (resp. $\mathfrak{F}_t$-sous-martingale) si l'égalité ci-dessus est remplacée par :

$$\mathbb{E}(M_t|\mathfrak{F}_s) \leq M_s \qquad (resp.\mathbb{E}(M_t|\mathfrak{F}_s) \geq M_s)$$

Par l'inégalité de Jensen, si M est une martingale alors M^2 est une sous-martingale puisque

$$E(M_t^2|\mathfrak{F}_s) \geqslant (E(M_t|\mathfrak{F}_s)^2 = M_s^2 \qquad \forall s \leqslant t$$

Une martingale peut être vue comme un processus d'erreurs, au sens où d'une part son espérance est constante (on pourra donc toujours supposer qu'elle est nulle) et d'autre part les incréments d'une martingale sont non corrélés :

$$\mathbf{cov}(M_t - M_s, M_v - M_u) = 0, 0 \leqslant s \leqslant t \leqslant u \leqslant v$$

Propriété 3.1 Le compensateur ou le processus de variation associée à une martingale M est l'unique processus croissant et prévisible $\langle M \rangle$ tel que

$$d\langle M \rangle = \mathbf{E}[(dM(t))^2 \mid \mathfrak{F}_{t^-}]$$

Propriété 3.2 Le processus de variation quadratique ou de variation optionnelle $[M]$ est la limite en probabilité de

$$\sum_i \{M(t_{i+1}) - M(t_i)\}^2$$

Afin de poursuivre la formalisation, il est nécessaire d'introduire une nouvelle définition :

Définition 3.1 *Un processus **prévisible** est une variable aléatoire mesurable définie sur l'espace produit* $([0, +\infty[\times \Omega, \mathcal{A})$ *muni de la tribu* $\mathcal{A}$ *engendrée par les ensembles de la forme* $]s,t] \times \Gamma$*, avec* $\Gamma \in \mathfrak{F}_s$.

La tribu des évènements prévisibles est engendrée par les processus adaptés à la filtration $(\mathfrak{F}_{t^-})_{t \geqslant 0}$ avec $\mathfrak{F}_{t^-} = \vee_{s<t} \mathfrak{F}_s$ et à trajectoires continues à gauche.
De manière intuitive, on peut dire qu'un processus prévisible est un processus dont la valeur en t est connue "juste avant" t. Ainsi un processus continu à gauche (et adapté) est prévisible du fait de la propriété de continuité.

Ces différents outils conduisent à la décomposition de *Doob-Meyer* d'un processus càd-làg adapté, qui exprime qu'un processus X càd-làg adapté est la différence de deux sous martingales (locales) si et seulement si il existe une unique décomposition de X sous la forme $X = A + M$ avec A un processus prévisible à variation bornée t M une martingale (locale) centré.

On en déduit en particulier que si M est une martingale, M^2 possède un compensateur prévisible, que l'on note $\langle M \rangle$.

3.2.1 Application aux modèles de durée

Définition 3.2 *Un processus* ***ponctuel*** $(N(t), t \geqslant 0)$ *est un processus à valeurs entières adapté à une filtration* $(\mathfrak{F}_t)_{t\geqslant 0}$ *tel que*

- $N(0) = 0$
- $N(t) < \infty$ *presque sûrement.*

et tel que les trajectoires soient continues à droite, constantes par morceaux et ne présentent que des sauts d'amplitude $+1$.

Remarque 3.1 Les processus ponctuels sont à trajectoires positives et croissantes, donc à variation bornée, et on peut donc définir pour un processus adaptés $X(t)$ l'intégrale $\int_0^t X(u)dN(u)$ comme une intégrale de *Stieljes* définie en **??**, trajectoire par trajectoire.Par exemple, en présence de censure le processus d'évènements non censurés $N^1(t) = 1_{\{T\leqslant t, D=1\}}$ peut s'écrire :

$$N^1(t) = \int_0^t C(u)dN(u)$$

avec $C(u) = 1_{[0,C]}(u)$. La censure agit donc comme un filtre. Comme un processus ponctuel est une sous martingale (puisqu'il est croissant), on lui associe son compensateur prèvisible , qui est croissant, de sorte que la différence entre le processus ponctuel et son compensateur soit une martingale.
De manière plus formelle on a le résultat suivant :

Proposition 3.1 *Si un processus ponctuel* $(N(t), t \geqslant 0)$ *adapté à* $\mathfrak{F}_t$ *est tel que* $\boldsymbol{E}[N(t)] < \infty$*, alors il existe un processus croissant continu à droite* Λ *tel que*

- $\Lambda(0) = 0$ *;*
- $\boldsymbol{E}[\Lambda(t)] < \infty$ *;*
- $M(t) = N(t) - \Lambda(t)$ *est une martingale.*

Lorsque Λ peut se mettre sous la forme $\int_0^t \lambda(u)du$, le processus λ s'appelle l'intensité du processus ponctuel.
La décomposition $N(t) = \Lambda(t) + M(t)$ exprime que le processus N peut se lire comme "observations= modèle+terme d'erreur" . On a en particulier $\mathbf{E}(N_t) = \mathbf{E}(\Lambda_t)$.
On cherche à déterminer le compensateur prévisible du processus $N(t) = 1_{\{X\leqslant t\}}$. On note :

$$N(t^-) = \lim_{u\to t} N(t)$$

et on s'intéresse à la loi de la variable $P(dN_t = 1 \mid N(t^-))$ avec $dNt = N(t+dt) - N(t)$, dN_t ne peut prendre que les valeurs 0 et 1. Par définition de h, on a :

$$P(dN_t = 1 \mid N(t^-) = 0) = h(t)dt$$

$$P(dN_t = 1 \mid N(t^-) = 1) = 0$$

que l'on peut aussi écrire :

$$\begin{aligned} P(dN_t = 1 \mid N(t^-)) &= h(t)dt \qquad \text{avec la probabilité } S(t) \\ &= 0 \qquad \text{avec la probabilité } 1 - S(t) \end{aligned}$$

On pose alors :

$$\lambda(t) = h(t)1_{\{X \geqslant t\}}$$

et $Y(t) = 1_{\{X \geqslant t\}}$, l'indicatrice de présence juste avant t.
Le processus $\lambda(t)$ est prévisible et $Y(t) = 1$ est équivalent à $N(t^-) = 0$, donc :

$$P(dN_t = 1 \mid N(t^-)) = \lambda(t)dt$$

De manière équivalente :

$$\mathbf{E}(dN_t \mid N(t^-)) = \lambda(t)dt$$

Les remarques ci-dessus impliquent que :

$$\begin{aligned} M(t) &= N(t) - \int_0^t \lambda(u)du \\ &= N(t) - \int_0^t h(u)Y(u)du \\ &= N(t) - H(t \wedge T). \end{aligned}$$

est une martingale centrée puisque $\mathbf{E}(dM_t \mid N(t^-) = 0$, et que l'intensité de processus N peut se calculer selon :

$$\lambda(t) = \lim_{u \to 0^+} \frac{1}{u} P[N(t+u) - N(t) = 1 \mid \mathfrak{F}_{t^-}]$$

Le processus $\lambda(t)$ est donc l'intensité de processus $N(t)$ qui est aléatoire. Conditionnellement au passé immédiat, l'accroissement de $N(t)$ entre t et $t + dt$ suit donc une loi de Bernoulli de paramètre $\lambda(t)dt$.

Le processus prévisible du processus d'évènements non censurés $N^1(t) = 1_{\{T \leqslant t, D=1\}}$ s'écrit :

$$\Lambda^1(t) = \int_0^t R(u)h(u)du.$$

avec $R(t) = 1_{\{T \leqslant t\}}$ l'indicatrice de présence à risque avant t.
On est donc passé du modèle statistique où l'on se donnait le couple (T, D) comme informations observées au moèle composé de (N^1, R).

Notations :

Dans le cas d'une population, dont on suppose que tous les individus ont la même fonction de hasard h, on associe à chaque individu :
un processus d'évènements non censurés N^{i1}

$$N_i^1(t) = 1_{\{T_i \leqslant t, D_i = 1\}}$$

et un indicatrice de présence sous risque $R_i(t)$:

$$R_i(t) = 1_{\{T_i \leqslant t\}}$$

et on construit les processus agrégés :

$$\overline{R}(t) = \sum_{i=1}^{n} R_i(t)$$

et

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

Le processus $\overline{N}^1(t)$ possède une intensité qui se met sous la forme :

$$\lambda(t) = \overline{R}(t)h(t)$$

avec $\overline{R}$ un processus prévisible et h la fonction de hasard, inconnue à estimer.

Dans la suite on note F la fonction de répartition du modèle non censuré, G la fonction de répartition de la censure et $T = X \wedge C$ la variable censurée. on note également :

$$S_0(t) = P(T > t, D = 0), S_1(t) = P(T > t, D = 1) \text{ et}$$
$$S_c(t) = S_0(t) + S_1(t) = (1 - F(t))(1 - G(t))$$

Ces processus vont permettre d'introduire simplement les estimateurs non paramétrique usuels.

3.3 Les estimateurs non paramétrique dans les modèles de durée

On notera en préambule que la distribution peut être, comme on l'a vu, caractérisée par différentes fonctions : fonction de hasard, fonction de hasard cumulée, fonction de répartition, densité ... Il est évident que l'estimation de la fonction de hasard est du même degré de complexité que l'estimation de la densité ; on se tournera donc de manière privilégiée vers l'estimation empirique du hasard cumulé ou de la fonction de survie, a priori plus simple. L'estimation de la fonction de hasard nécessitera alors de régulariser l'estimateur de la fonction de hasard cumulée, qui sera en général discontinu.

Les deux estimateurs principaux dans ce contexte sont l'estimateur de **Nelson-Aalen** du taux de hasard cumulé et l'estimateur de **Kaplan-Meier** de la fonction de survie.

3.3.1 L'estimateur de Nelson Aalen du taux de hasard cumulé

3.3.1.1 Présentation générale

Le fait que $M(t) = \overline{N}^1(t) - \int_0^t \overline{R}(u)h(u)du$ soit une martingale centrée suggère de proposer $\overline{N}^1(t)$ comme estimateur de $\int_0^t \overline{R}(u)h(u)du$. Mais alors le processus $\int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u) > 0} dM(u)$

est également une martingale et on a par construction de M :

$$\int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)>0} dM(u) = \int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)>0} d\overline{N}^1(t) - \int_0^t h(u)du = \int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)>0} d\overline{N}^1(t) - H(t)$$

pour autant que t soit tel que $\overline{R}(t) > 0$.Ainsi

$$\hat{H}(t) = \int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)>0} d\overline{N}^1(t)$$

est un estimateur de H. Cet estimateur s'appelle de Nelson-Aalen. il a été proposé initialement par Nelson[1972] [19].

On peut en donner une autre justification, en remarquant que la fonction de hasard cumulé vérifie, par construction :

$$H(u+du) - H(u) \approx h(u)du$$

et $h(u)du = P(\text{sortie entre } u \text{ et } u+du \mid \text{en vie en } u)$; un estimateur naturel de cette quantité est donc $\dfrac{\overline{N}^1(u+du) - \overline{N}^1(u)}{\overline{R}(u)} = \dfrac{d\overline{N}^1(u)}{\overline{R}(u)}$ si $\overline{R}(u) > 0$, de sorte qu'en sommant sur un découpage de$[0,t]$ suffisamment fin pour chaque subdivision contienne au plus un saut on obtient :

$$\hat{H}(t) = \int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)>0} d\overline{N}^1(t)$$

ce qui est bien l'expression précédente. Comme les processus considérés ici sont purement à sauts on peut, en notant $\triangle\overline{N}(t) = \overline{N}(t) - \overline{N}(t^-)$, mettre cette expression sous la forme :

$$\hat{H}(t) = \sum_{\{i<T_i<t\}} \frac{\triangle\overline{N}(T_i)}{\overline{R}(T_i)}$$

En posant $d(t) = \triangle\overline{N}(t)$ le nombre de décès en t et $r(t) = \overline{R}(t)$ l'effectif sous risque juste avant t, on peut ainsi réécrire l'équation ci-dessus sous la forme intuitive suivante :

$$\hat{H}(t) = \sum_{\{i<T_i<t\}} \frac{d(T_i)}{r(T_i)} = \sum_{T_i<t} \frac{d_i}{n-i+1},$$

la seconde égalité n'étant vrai que si il n'y a pas d'ex-aequo. La fonction $\hat{H}$ est continue à droite. On peut vérifier que cet estimateur est biaisé et sous-estime en moyenne la fonction de hasard cumulée. En effet,

$$\hat{H}(t) = \int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)} d\overline{N}^1(u) = \int_0^t \frac{1_{\{\overline{R}(u)>0\}}}{\overline{R}(u)} (dM(u) + \overline{R}(u)h(u)du)$$

Comme M est une martingale, il vient en prenant l'espérance des deux membres de l'équation ci-dessus $E\left[\hat{H}(t)\right] = \int_0^t E\left(1_{\{\overline{R}(u)>0\}}\right) h(u)du$. Mais :

$$E\left[1_{\{\overline{R}(u)>0\}}\right] = P\left[\overline{R}(u)>0\right] = 1 - P\left[\overline{R}(u)=0\right].$$

On en déduit finalement :

$$E\left[\hat{H}(t)\right] = \int_0^t h(u)du - \int_0^t P\left[\overline{R}(u)=0\right]h(u)du = H(t) - \int_0^t P\left[\overline{R}(u)=0\right]h(u)du$$

ce qui implique que $E\left[\hat{H}(t)\right] \leqslant H(t)$: l'estimateur de Nelson-Aalen a bien tendance à sous-estimer la fonction de hasard cumulée du modèle.

3.3.1.2 Variance de l'estimateur de Nelson-Aalen

Il résulte de l'approximation effectuée à la section précédente que l'accroissement du processus $\overline{N}^1(t)$ entre t et $t+u$ suit approximativement une loi de Poisson de paramètre

$$\int_t^{t+u} \overline{R}(s)h(s)ds \approx \overline{R}(t)h(t)u.$$

En effet, on avait vu que conditionnellement au "passé immédiat", l'accroissement de $N^1(t)$ entre t et $t+dt$ suit donc une loi de Bernoulli de paramètre $h(t)R(t)dt$. La somme sur les différents individus conduit donc à une variable binomiale, que l'on peut approcher par une loi de Poisson en choisissant $dt = \dfrac{u}{n}$. On en déduit donc que, conditionnellement à $\overline{R}(t)$,

$$V\left(\frac{\overline{N}^1(t+u)-\overline{N}^1(t)}{\overline{R}(t)}\right) \approx \frac{h(t)u}{\overline{R}(t)}$$

or on a vu à la sélection précédente que $h(t)u$ pouvait être estimer par $\dfrac{\overline{N}^1(t+u)-\overline{N}^1(t)}{\overline{R}(t)}$, d'où l'estimateur de la variance

$$\hat{V}\left(\frac{\overline{N}^1(t+u)-\overline{N}^1(t)}{\overline{R}(t)}\right) \approx \frac{\overline{N}^1(t+u)-\overline{N}^1(t)}{\overline{R}(t)^2}$$

ce qui conduit finalement à proposer comme estimateur de la variance de $\hat{H}$:

$$\hat{V}(\hat{H}(t)) = \sum_{\{i/T_i \leqslant t\}} \frac{\triangle \overline{N}^1(T_i)}{\overline{R}(T_i)^2}$$

qui peut s'écrire avec les notations simplifiées, en l'absence d'exaequo :

$$\hat{V}(\hat{H}(t)) = \sum_{\{i/T_i \leqslant t\}} \frac{d(T_i)}{(n-i+1)^2}$$

3.3.1.3 Propriétés asymptotiques

L'estimateur de Nelson-Aalen est asymptotiquement gaussien ; plus précisément on a le résultat suivant : Si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sqrt{n}(\hat{H}-H) \longrightarrow W_H$$

avec W_H un processus gaussien centré de covariance

$$\rho(s,t) = \int_0^{s\wedge t} \frac{dS_1(u)}{S_c(u)^2}$$

avec $S_c(t) = (1-F(t))(1-G(t))$ et $S_1(t) = P(T > t, D = 1)$.

3.3.2 L'estimateur de Kaplan-Meier de la fonction de survie

On peut remarquer que l'estimateur de Nelson-Aalen du taux de hasard cumulé conduit à un estimateur naturel de la fonction de survie, en exploitant la relation $S(t) = \exp(-H(t))$; on peut ainsi proposer comme estimateur de la fonction de survie

$$\hat{S}(t) = \exp(-\hat{H}(t)).$$

Cet estimateur est l'estimateur de Harrington et Fleming ; sa variance peut être obtenue par la méthode Delta qui, sous des conditions raisonnables de régularité de la fonction f permet d'écrire que $V(f(X)) \approx \left(\frac{df}{dx}(E(X))\right)^2 V(X)$. En effet, si une variable aléatoire X est proche de $\mu + \sigma Z$ avec σ petit et Z centrée réduite, on remarque que pour une fonction $x \to f(x)$ suffisamment régulière, en effectuant le développement limité

$$f(\mu + h) \approx f(\mu) + h\frac{df}{dx}(\mu)$$

On trouve que

$$V(f(X)) \approx V\left(f(\mu) + \sigma Z \frac{df}{dx}(\mu)\right) = \sigma^2 \frac{df}{dx}(\mu)^2$$

En prenant ici $f(x) = e^{-x}$, on trouve que

$$V(\hat{S}) \approx e^{-2E(\hat{H})} V(\hat{H}) \approx \hat{S}^2 V(\hat{H})$$

Ce qui conduit à l'estimateur de la variance :

$$\hat{V}(\hat{S}) = \exp\left(-2 \sum_{\{i/t_i \leqslant t\}} \frac{d(t_i)}{n-i+1}\right) \sum_{\{i/t_i \leqslant t\}} \frac{d(t_i)}{(n-i+1)^2}$$

Toutefois, cet estimateur peut être amélioré, ce qui amène à introduire l'estimateur de Kaplan-Meier.

3.3.2.1 Présentation générale

L'estimateur de Kaplan-Meier (KAPLAN et MEIER [1958] [17]) peut également être introduit via les processus ponctuels, en remarquant que la fonction de survie de base du modèle est l'unique solution de l'équation intégrale suivante :

$$S(t) = 1 - \int_0^t S(u^-) h(u) du.$$

L'équation ci-dessus exprime simplement le fait que la somme des survivants en t et des individus sortis avant t est constante. Lorsque la fonction de survie est continue, la démonstration est immédiate en effectuant le changement de variable $\nu = \ln S(u)$, $d\nu = -h(u)du$.

En remplaçant $h(u)du$ par son estimateur $\dfrac{d\overline{N}^1(u)}{\overline{R}(u)}$ introduit à la section précédente on peut proposer un estimateur de la fonction de survie en cherchant une solution à l'équation :

$$\hat{S}(t) = 1 - \int_0^t \hat{S}(u^-)\frac{d\overline{N}^1(u)}{\overline{R}(u)}$$

On peut montrer qu'il existe une unique solution à cette èquation, et on obtient alors l'estimateur de Kaplan-Meier de la fonction de survie. Cet estimateur peut s'exprimer à l'aide de l'estimateur de Nelson-Aalen de la manière suivante :

$$\hat{S}(t) = \prod_{s \leqslant t} \left(1 - \triangle \hat{H}(s)\right)$$

où $\triangle \hat{H}(s) = \hat{H}(s) - \hat{H}(s^-)$. On peut toutefois proposer une construction explicite plus intuitive de cet estimateur.

La construction heuristique de l'estimateur de Kaplan-Meier s'appuie sur la remarque suivante :

la probabilité de survivre au delà de $t > s$ peut s'écrire :

$$S(t) = P(T > t \mid T > s)P(T > s) = P(T > t \mid T > s)S(s)$$

On peut renouveler l'opération ce qui fait apparaître des produits de termes en $P(T > t \mid T > s)$; si on choisit comme instants de conditionnement les instants où se produit un événement(sortie ou censure), on se ramène à estimer des probabilités de la forme :

$$p_i = P\left(T > T_{(i)} \mid T > T_{(i-1)}\right)$$

p_i est la probabilité de survivre sur l'intervalle $]T_{(i-1)}, T_{(i)}]$ sachant qu'on était vivant à l'instant $T_{(i-1)}$. Un estimateur naturel de $q_i = 1 - p_i$ est $\hat{q}_i = \dfrac{d_i}{r_i} = \dfrac{d_i}{n-i+1}$.

On observe alors qu'à l'instant $T_{(i)}$, et en l'absence d'ex aequo, si $D_{(i)} = 1$ alors il y a sortie par décès donc $d_i = 1$, et dans le cas contraire l'observation est censurée et $d_i = 0$. L'estimateur de Kaplan-Meier s'écrit donc finalement :

$$\hat{S}(t) = \prod_{T_{(i)} < t} \left(1 - \frac{1}{n-i+1}\right)^{D_{(i)}}$$

En pratique cependant on est confronté à la présence d'ex aequo ; on suppose alors par convention que les observations non censurées précèdent toujours les observations censurées. On obtient l'expression suivante de l'estimateur :

$$\hat{S}(t) = \prod_{T_{(i)} < t} \left(1 - \frac{d_i}{r_i}\right)$$

L'estimateur $\hat{S}(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier est un estimateur de maximum de vraisemblance.$\hat{S}(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car les temps de décès ne sont pas connus).

3.3.2.2 Variance de l'estimateur de Kaplan-Meier

L'estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier est

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{i:T_{(i)}\leqslant t} \frac{d_i}{r_i(r_i - d_i)}$$

Il est obtenu en utilisant l'approximation suivante :

$$\hat{V}\left(\ln \hat{S}(t)\right) \approx \sum_{i:T_{(i)}\leqslant t} \frac{d_i}{r_i(r_i - d_i)}$$

et en appliquant la méthode Delta $V(f(X)) \approx \left(\frac{df}{dx}(E(X))\right)^2 V(X)$ pour montrer que

$$\hat{V}\left(\ln \hat{S}(t)\right) \approx \frac{1}{\hat{S}(t)^2} V(\hat{S}(t)).$$

3.3.2.3 Principales propriétés

L'estimateur de Kaplan-Meier possède un certain nombre de bonnes propriétés qui en font la généralisation naturelle de l'estimateur empirique de la fonction de répartition en présence de censure : il est convergent, asymptotiquement gaussien, cohérent et est également un estimateur du maximum de vraisemblance généralisé. Toutefois, cet estimateur est biaisé positivement. La cohérence de l'estimateur signifie que la propriété suivante est vérifiée :

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i > t\}} + \sum_{i=1}^{n} 1_{\{T_i > t, D_i = 0\}} \frac{\hat{S}(t)}{\hat{S}(T_i)}$$

Cette formule signifie que les survivants au-delà de t sont la somme :

- des individus ni morts ni censurés avant t ;
- des individus qui, censurés en T_i avant t, survivent après t avec la probabilité conditionnelle $\frac{\hat{S}(t)}{\hat{S}(T_i)}$.

3.3.2.4 Propriétés asymptotiques

L'estimateur de Kaplan-Meier est asymptotiquement gaussien ; précisément on a le résultat suivant :

Proposition 3.2 si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sqrt{n}(\hat{S} - S) \longrightarrow W_s$$

avec W_s un processus gaussien centré de covariance :

$$\rho(s,t) = S(s)S(t) \int_0^{s \wedge t} \frac{dF(u)}{(1-F(u))^2(1-G(u))}$$

En particulier lorsque le modèle n'est pas censuré (ie $G(u) = 0$) on retrouve le résultat classique présenté en ci-dessus. L'intérêt de résultats de convergence au niveau du processus lui même plutôt que pour un instant fixé est que l'on peut en déduire des bandes de confiance asymptotique pour l'estimateur de Kaplan-Meier.

3.3.3 Estimation de la survie par la méthode actuarielle

La méthode actuarielle repose sur le même principe de construction que l'estimateur de Kaplan-Meier. La différence est que les probabilités conditionnelles sont estimées sur des intervalles fixés par l'utilisateur et non déterminés par les temps d'événements. Ces intervalles sont généralement de longueur égale, par exemple, un mois, un trimestre, une année.
Considérons k intervalles de temps $[0,t_1[, [t_1,t_2[, ..., [t_{k-1},\infty[$, fixés à priori. Définissons,

- d_i le nombre de décès dans le i^{eme} intervalle $[t_{i-1},t_i[$ (avec $t_0 = 0$ et $t_k = \infty$),
- n_{i-1} le nombre de sujets vivant au temps t_{i-1},
- c_i le nombre de sujets censurés dans l'intervalle $[t_{i-1},t_i[$,
- r_i le nombre de sujets à risque dans l'intervalle $[t_{i-1},t_i[$,

Afin de simplifier les calculs, on suppose généralement que les censures sont réparties uniformément dans l'intervalle, c'est-à-dire, que les sujets censurés sont exposés en moyenne un demi-intervalle. Dans le calcul des individus à risque, leur contribution pour l'intervalle $[t_{i-1},t_i[$ est donc $c_i/2$. Le nombre d'individus à risque pour l'intervalle $[t_{i-1},t_i[$ est donc

$$r_i = n_{i-1} - \frac{c_i}{2}$$

Alors la probabilité $p_i = P(X \leqslant t_i \mid X > t_{i-1})$ de mourir dans l'intervalle $[t_{i-1},t_i[$ sachant que l'on était vivant en t_{i-1} est estimée par

$$\hat{p}_i = \frac{d_i}{r_i}$$

L'estimateur de la fonction de survie est donc

$$\hat{S}(t) = \prod_{i:t_i \leqslant t} \left(1 - \frac{d_i}{r_i}\right)$$

La formule de Greenwood permet d'obtenir une estimation de la variance ,

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \leqslant t} \frac{d_i}{r_i(r_i - d_i)}.$$

3.4 Prise en compte de variables explicative

Lorsque la population étudiée est hétérogène, il est important de prendre en compte les spécificités de chaque sous-groupe. En supposant que l'hétérogénéité est la conséquence d'un mélange de sous-populations caractérisées chacune par des variables observables, on s'intéresse ici à des modélisations de la fonction de hasard intégrant l'effet des variables explicatives. Cette question a déjà été abordée dans un contexte paramétrique et semiparamétrique (modèle de Cox), on s'intéresse ici au cas non paramétrique.

3.4.1 Le modèle additif d'Aalen

La fonction de hasard est supposée s'écrire :

$$h(t) = X^T(t)\beta(t)$$

avec $X^T(t) = (X_1(t), ..., X_p(t))$ un vecteur de variables explicatives et $\beta(t)$ un processus p-dimensionnel localement intégrable. On peut de manière équivalente dire que l'inensité du modèle de comptage sous-jacent s'écrit :

$$\lambda(t) = R(t)X^T(t)\beta(t)$$

On dispose d'un ensemble d'observations $(N_i^1(t), R_i(t), X^i(t))_{1 \leqslant i \leqslant n}$ et on cherche à estimer le vecteur $\beta(t)$; en pratique on va être en mesure de construire un estimateur de $B(t) = \int_0^t B(u)du$ en s'appuyant sur les remarques qui suivent. On note :

$$\lambda(t) = (\lambda_1(t), ..., \lambda_n(t))^T$$

et

$$N^1(t) = (N_1^1(t), ..., N_n^1(t))^T$$

puis

$$X(t) = R_1(t)X^1(t), ..., R_n(t)X^n(t))^T$$

qui est une matrice de taille $n \times p$. Avec ces notations on a en désignant par $\Lambda(t) = \int_0^t \lambda(u)du$ le processus vectoriel de taille n des intensités cumulées, $M(t) = N^1(t) - \Lambda(t)$ est une martingale. En observant alors que :

$$dN^1(t) = X(t)\beta(t)dt + dM(t) = X(t)dB(t) + dM(t)$$

comme le terme $dM(t)$ est centré et que les incréments de la martingale sont non corrélés, on peut chercher à estimer les incréments $dB(t)$ par des techniques classiques de régression linéaire.

Chapitre 4

Modèles semi paramétriques

Sommaire

4.1 Introduction

Si on ne peut pas spécifier entièrement la famille de loi à la quelle appartient la durée de vie, ou bien si l'effet relatif des diverses covariables sur le phénomène étudié est pour nous le plus important à étudier, il est souvent fructueux d'utiliser les modèles semi-paramétrique. Ces derniers n'introduisent pas d'hypothèses (autre que régularité) sur les fonctions de densité et/ou de hasard, mais font des hypothèses sur la manière dont les diverses covariables vont influencer le déroulement du phénomène temporel. On distingue deux grandes classes de modèles.

4.2 Les modèles à hasard proportionnels

Ces modèles expriment un effet multiplicatif des diverses covariables sur la fonction de hasard (modèle à structure multiplicative). On introduit une fonction de hasard de base qui donne la forme générale du hasard et qui est commune à tous les individus. Les modèles à hasards proportionnels se caractérisent par la relation suivante, pour tout $t > 0$;

$$h(t \mid Z) = h_0(t)g(\beta, Z),$$

où Z est un vecteur de covariables, β le paramètre d'intérêt et h une fonction positive.
La fonction de hasard est le produit d'une fonction qui ne dépend que du temps et d'une fonction qui n'en dépend pas. En général, on suppose que l'effet des covariables se résume à une quantité réelle $\beta' Z$, c'est-à-dire

$$h(t \mid Z) = h_0(t)g(\beta' Z).$$

Ce modèle est dit à risques proportionnels car, quels que soient deux individus i et j qui ont pour covariables Z_i et Z_j ; le rapport des fonctions de hasard ne varie pas au cours du temps,

$$\frac{h(t \mid Z_i)}{h(t \mid Z_j)} = \frac{g(\beta' Z_i)}{g(\beta' Z_j)}.$$

Les fonctions de hasard sont donc proportionnelles. C'est une conséquence du modèle mais c'est aussi une hypothèse qu'il faudra vérifier. Le rapport des fonctions de hasard est par définition un risque relatif à l'instant t des sujets de caractéristiques Z_i par rapport aux sujets de caractéristiques Z_j . Un cas particulier très important est le modèle de Cox, qui suppose que la fonction g est la fonction exponentielle, c'est-à-dire,

$$h(t \mid Z) = h_0(t) \exp(\beta' Z)$$

D'autres choix de fonctions g sont possibles, néanmoins la fonction exponentielle est très souvent utilisée dans la littérature car ses valeurs sont toujours positives et $\exp(0) = 1$.

Remarque 4.1 *Si h_o et /ou g ont une forme inconnue le modèle se dit semi-paramérique.*

4.3 Le modèle semi-paramétrique de COX

Nous commençons par définir les notations utilisées dans cette section, donner la formulation du modèle de régression de Cox ainsi que la méthode du maximum de vraisemblance partielle permettant d'estimer le paramètre d'intérêt du modèle.

Le modèle de régression de Cox fait partie de la famille des modèles à risques multiplicatifs et a été introduit par Cox (1972)[10]. C'est *un modèle semi-paramétrique* qui permet de ne spécifier que la modélisation relative à l'influence des covariables via un vecteur de régression et d'éviter ainsi le choix parfois difficile d'un modèle totalement paramétrisé. Il a été décrit originellement par la formulation de la fonction de risque instantané de la donnée de survie.

4.3.1 Définitions et notations

Considérons un échantillon de n individus.
Soit τ la date de point et β le paramétre de régression dont la dimension est égale à p.
Soient, pour $i \in \{1,\dots,n\}$,

- X_i la date de survenue de l'évènement chez l'individu i ;
- C_i la date de censure correspondant ;
- $T_i = X_i \wedge C_i$;
- $\delta_i = 1_{\{X_i \leqslant C_i\}}$;
- $Z_i = Z_{1i},...,Z_{pi}$ le vecteur de dimension p des covariables ;
- $Y_i(t) = 1_{\{T_i \geqslant t\}}$

Soient aussi

- $N(t) = \{N_i(t), 0 \leqslant t \leqslant \tau, i = 1,\dots,n\}$ le processus de comptage multivarié pour les n individus ;
- $\Lambda(t) = \{\Lambda_i(t), 0 \leqslant t \leqslant \tau, i = 1,\dots,n\}$ le compensateur de $N(t)$ par rapport à la filtration $\mathcal{F}_t$.

Le **modèle de Cox (1972)** [10] spécifie que le risque instantané s'écrit :

$$h_i(t) = h_0(t)\exp(\beta^t Z_i),$$

Où $h_0(t)$ est **la fonction de risque de base**.
Il s'agit d'un **modèle semi-paramétrique à risques proportionnels** :

- semi-paramétrique, du fait de la présence, dans la définition du risque instantané, d'une partie paramétrique (la partie de régression $\exp(\beta^t Z_i)$ et d'une partie non-paramétrique (le risque de base $h_0(t)$) ;
- à risques proportionnels, car quels que soient i et j, le rapport des risques instan-

tanés de deux individus ne varie pas au cours du temps :

$$\frac{h_i(t)}{h_j(t)} = \exp[\beta^t(Z_i - Z_j)]$$

4.3.2 Estimation

4.3.2.1 Présentation

Nous avons présenté,dans la section 1.7 , la théorie mathématique des processus de comptage sur laquelle vont reposer les résultats qui suivent. Ainsi, les définitions de la vraisemblance dans des différents cas de modèles dans 1.6 nous assurent que la vraisemblance complète associée à un processus ponctuel N^* simple − i.e. non filtré , c.-à-d. non censuré − est de la forme

$$\mathcal{L}^*(\beta) = \prod_{t\leq\tau}\prod_{i=1}^{n}\left\{(h_i(t)Y_i(t))^{\Delta N_i^*(t)}\right\} \times \exp\left[-\sum_i Y_i(\tau)A_i(\tau)\right]. \tag{4.1}$$

Nous passons de la vraisemblance associée au processus simple N^* − dite ***vraisemblance complète*** − à celle associée au processus censuré N dite − ***vraisemblance partielle*** − en supprimant, dans 4.1, les termes correspondant, pour l'intervalle de temps dt, à la contribution du processus indicateur prévisible de censure C .
Ainsi, la vraisemblance partielle dans le cadre du modèle de Cox s'écrit

$$\begin{aligned}\mathcal{L}(\beta) &= \prod_{t\leq\tau}\prod_{i=1}^{n}\left\{(h_i(t)Y_i(t))^{\Delta N_i^*(t)}\right\} \times \exp\left[-\sum_i Y_i(\tau)A_i(\tau)\right] \\ &= \prod_t\prod_i\left\{\left[h_0(t)Y_i(t)\exp(\beta^t Z_i)\right]^{\Delta N_i^*(t)}\right\} \times \exp\left[-\int_0^\tau S^{(0)}(\beta,u)h_0(u)du\right]\end{aligned} \tag{4.2}$$

avec

$$S^{(0)}(\beta,t) = \sum_j Y_j(t)\exp(\beta^t Z_j).$$

À β fixé, la maximisation de 4.2 suivant $\Delta A_0(t)$ conduit à

$$\Delta\widehat{A}_0(t) = \frac{\Delta N_+(t)}{S^{(0)}(\beta,t)}.$$

Par conséquent, toujours à β fixé, on estime $A_0(t)$ par

$$\widehat{A}_0(t) = \int_0^t \frac{J(u)}{S^{(0)}(\beta,u)} dN_+(u). \tag{4.3}$$

Avec $J(u) = 1_{\{Y_1(u)+\cdots+Y_n(u)>0\}}$
$\widehat{A}_0(t)$ est appelé **estimateur de Breslow** (Breslow, 1974) [8].

En remplaçant, dans 4.2, $A_0(t)$ par son estimation obtenue en 4.3, nous obtenons pour expression de la vraisemblance partielle

$$\begin{aligned}
\mathcal{L}(\beta) &= \prod_t \prod_i \left\{ \left(d\widehat{A}_0(t) Y_i(t) \exp(\beta^t Z_i) \right) \right\} \times \exp\left[- \int_0^\tau S^{(0)}(\beta, u) d\widehat{A}_0(t) \right] \\
&= \prod_t \prod_i \left\{ (Y_i(t) \exp(\beta^t Z_i))^{\Delta N_i(t)} (d\widehat{A}_0(t))^{\Delta N_i(t)} \right\} \times \exp\left[- \int_0^\tau S^{(0)}(\beta, u) \frac{J(u) dN_+(u)}{S^{(0)}(\beta, u)} \right] \\
&= \prod_t \prod_i \left\{ [Y_i(t) \exp(\beta^t Z_i)]^{\Delta N_i(t)} \times \left[\frac{J(t) dN_+(t)}{S^{(0)}(\beta, u)} \right]^{\Delta N_i(t)} \times \exp\left[- \int_0^\tau J(u) dN_+(u) \right] \right. \\
&= \prod_t \prod_i \left\{ \left[\frac{Y_i(t) \exp(\beta^t Z_i)}{S_0(\beta, t)} \right]^{\Delta N_i(t)} \right\} \times \exp\left[- \int_0^\tau S^{(0)}(\beta, u) \frac{J(u) dN_+(u)}{S^{(0)}(\beta, u)} \right] \\
&= L(\beta) \times \prod_t \prod_i \{ J(t) dN_+(t) \}^{\Delta N_i(t)} \times \exp\left[- \int_0^\tau J(u) dN_+(u) \right].
\end{aligned}$$

avec

$$L(\beta) = \prod_t \prod_i \left(\frac{Y_i(t) \exp(\beta^t Z_i)}{S_0(\beta, t)} \right)^{\Delta N_i(t)} \tag{4.4}$$

dépendant de β (le reste de la vraisemblance étant indépendant de β).
Par définition, $L(\beta)$ est **la vraisemblance partielle de Cox**.

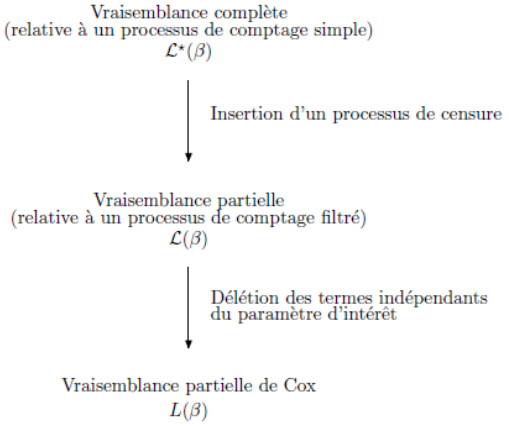

FIGURE 4.1 – Les vraisemblances successives.

Remarque 4.2 La présence, dans 4.4, des deux produits (l'un suivant t, l'autre suivant i) découle de la formalité de l'écriture mathématique. Ainsi, ces deux signes peuvent paraître redondant, dans la mesure où l'hypothèse d'absence d'ex-aequo entraîne que chaque processus de comptage $N_i(t)$ ne vaut 1 que pour une et une seule valeur de t.

Considérons maintenant la fonction de log-vraisemblance partielle de Cox sur l'intervalle $[0,t[$:

$$\log L(\beta,t) = \sum_i \int_0^\tau \left[\beta^t Z_i - \log S^{(0)}(\beta,u)\right] dN_i(u). \tag{4.5}$$

Le vecteur score

$$\mathbf{U}(\beta,t) = \frac{\partial \log L(\beta,t)}{\partial \beta}$$

peut s'écrire

$$\begin{aligned} \mathbf{U}(\beta,t) &= \sum_i \int_0^t [Z_i - E(\beta,u)]\, dN_i(u) \\ &= \sum_i \int_0^t [Z_i - E(\beta,u)]\, dM_i(u) \end{aligned} \tag{4.6}$$

Où

$$E(\beta,u) = \frac{S^1(\beta,t)}{S^0(\beta,t)} \tag{4.7}$$

et

$$S^1(\beta,t) = \sum_i Y_i(t) \exp(\beta^t Z_i) Z_i.$$

$Z_i - E(\beta,u)$ étant un vecteur de processus prévisible, $\mathbf{U}(\beta,t)$ est une somme de n martingales vectorielles, et est donc lui-même une martingale.

La suite de martingales $M^{(n)}(t) = n^{-1/2}\mathbf{U}(\beta,t)$ vérifie les conditions d'application du théorème de Rebolledo. En appliquant ce dernier, nous déduisons de la loi du processus limite $M^{(\infty)} = M^{(\tau)}$ le résultat suivant :

Proposition 4.1 *Soit $\widehat{\beta}$ l'estimateur du maximum de vraisemblance partielle de cox, i.e. la quantité vérifiant*

$$\mathbf{U}(\beta,\tau) = 0 \tag{4.8}$$

Alors

$$\widehat{\beta} \longrightarrow \mathcal{N}(\beta_0, \mathfrak{I}^{-1}(\widehat{\beta}),$$

*Où $\mathfrak{I}(\beta)$ est la **matrice d'information de Fisher** :*

$$\begin{aligned} \mathfrak{I}(\beta) &= -\frac{\partial^2 L(\beta)}{\partial \beta^2} \\ &= -\sum_{i=1}^n \int_0^\tau \left\{ \frac{S^2(\beta,s)}{S^0(\beta,s)} - E(\beta,s)^{\otimes 2} \right\} dN_i(s) \end{aligned}$$

avec

$$S^2(\beta,s) = \sum_{i=1} Y_i(s) \exp(\beta^t Z_i) Z_i^{\otimes 2}.$$

$\mathfrak{I}^{-1}(\beta)$ inverse de la matrice d'information de Fisher, fournit une estimation de la variance de $\widehat{\beta}$.

4.3.2.2 Résolution numérique

Pour rèsoudre l'équation du score 4.8, l'**algorithme de Newton-Raphson** est habituellement employé.

Partant d'une solution initiale $\widehat{\beta}_0 = 0$, l'algorithme consiste en la succession d'itérations de la forme

$$\widehat{\beta}^{j+1} = \widehat{\beta}^{j} - \left[\frac{\partial^2 \log L(\widehat{\beta}^{j},\tau)}{\partial^2 \beta}\right]^{-1} \frac{\partial \log L(\widehat{\beta}^{j},\tau)}{\partial \beta}.$$

Le terme qui suit le signe moins est le pas itératif de l'algorithme de Newton-Raphson.

Si la fonction de vraisemblance évaluée en $\widehat{\beta}^{j+1}$ est inférieure à celle évaluée en $\widehat{\beta}^{j}$, alors $\widehat{\beta}^{j+1}$ est recalculé en utilisant, cette fois-ci, la moitié du pas itératif.

Ces étapes ne succèdent jusqu'à ce que la convergence soit obtenue, c'est-à-dire jusqu'à ce que $\widehat{\beta}^{m+1}$ soit suffisamment proche de $\widehat{\beta}^{m}$. L'estimateur du maximum de vraisemblance de β est alors $\widehat{\beta} = \widehat{\beta}^{m+1}$.

4.4 Tests

Trois tests de l'hypothèse nulle $H_0 : \beta = \beta_0$ peuvent être déduits du résultat concernant la convergence asymptotique de $\widehat{\beta}$.

4.4.1 Test du rapport de vraisemblance

Ce test, très couramment utilisé en statistique, découle d'un développement de Taylor à l'ordre 2 de $\log L(\beta)$, puis de propriétés de convergence en loi.

Il s'énonce comme suit :

$$2[\log L(\widehat{\beta}) - \log L(\beta_0)] \rightsquigarrow \chi^2(p).$$

Ce test mesure la différence des valeurs prises par le logarithme de la vraisemblance en $\widehat{\beta}$ et β_0 ; l'espérance de cette quantité doit être nulle sous H_0.

4.4.2 Test de Wald (ou du maximum de vraisemblance)

D'après le résultat de la proposition 4.1, nous avons

$$\sqrt{\mathfrak{I}(\widehat{\beta})}(\widehat{\beta} - \beta_0) \longrightarrow \mathcal{N}(0,1)$$

Or si une v.a.X p-dimensionnelle suit une loi normale centrée réduite, alors X^2 suit une loi du chi-deux à p degrés de liberté. Ainsi,

$$(\widehat{\beta} - \beta_0)^t \mathfrak{I}(\widehat{\beta})(\widehat{\beta} - \beta_0) \rightsquigarrow \chi^2(p).$$

Il mesure l'écart entre $\widehat{\beta}$ et β_0, qui est nul en moyenne sous H_0 car $\widehat{\beta}$ est asymptotiquement sans biais.

4.4.3 Test de score

Il est possible de montrer que

$$\frac{\partial \log L(\beta,t)}{\partial \beta} \mid_{\beta=\beta_0} \longrightarrow \mathcal{N}(0,\mathfrak{I})$$

En notant $\mathbf{U}(\beta,t) = (\partial \log L(\beta,t))/\partial\beta$, nous obtenons par conséquent

$$\left[\mathbf{U}(\beta,\tau)^t \mathfrak{I}^{-1}(\beta_0)\mathbf{U}(\beta,\tau)\right] \rightsquigarrow \chi^2(p).$$

Ce test mesure la pente de la tangente en β_0. Sous H_0, le maximum de vraisemblance est obtenu pour une valeur $\widehat{\beta}$ proche de β_0. La pente en β_0 diffère donc peu de 0, elle est nulle en moyenne sous H_0.

4.5 Critères d'adéquation au moèle de Cox

L'appréciation de la justesse de l'adéquation (*goodness of fit*) passe le plus souvent par une comparaison graphique des résidus des différents modèles que nous désirons comparer ; une alternative, employant certaines statistiques, peut quelquefois se présenter.

4.5.1 Dans sa globalitè

Définissons les **résidus de Cox-Snell** comme étant les quantités suivantes :

$$R_i = \widehat{A}_0(T_i)e^{\widehat{\beta}^t Z_i}.$$

Si le modèle est correct et si les paramètres sont proches de leurs vraies valeurs, les R_i doivent alors constituer un échantillon censuré de distribution exponentielle unitaire.

Par suite, la représentation graphique de l'estimateur de Nelson-Aalen du risque cumulé en fonction des R_i doit approcher la première bissectrice.

4.5.2 Concernant la forme fonctionnelle des covariables

Il est possible que la forme fonctionnelle des covariables, telle qu'elle est spécifiée par le modèle de Cox - soit $\exp(\beta^t Z)$ -, ne soit pas exacte. Considérons donc le modèle à deux covariables

$$\begin{aligned}\Lambda(t,Z,X) &= h(Z^{(1)})\exp(\beta^t Z^{(2)})\Lambda_0(t) \\ &= \exp\left(f(Z^{(1)}\right)\exp(\beta^t Z^{(2)}\Lambda_0(t))\end{aligned} \tag{4.9}$$

où la forme fonctionnelle pour $Z^{(2)}$ est connue, tandis que la fonction positive $h(Z^{(1)})$ est, elle, inconnue.

Définissons également les résidus martingales par

$$\widehat{M_i} = \delta_i - \widehat{A}_0(T_i)\exp(\beta^t Z_i^{(2)}),$$

où T_i est le temps d'observation concernant le i^e individu.

Dans le cas d'un modèle linéaire classique, les résidus sont la différence entre les valeurs observées et les valeurs attendues.

Leur représentation graphique en fonction de $Z^{(1)}$ permet d'obtenir des estimations de h ou de f.

Si l'on note $\widehat{M}$ le résidu martingale lorsque le modéle 4.9 est le modèle adéquat, mais que l'on a ignoré $Z^{(1)}$, et si $Z^{(1)}$ et $Z^{(2)}$sont indépendantes, alors (Therneau *et al.*, 1990)[23].

$$\mathbb{E}\left[\widehat{M}(t) \mid Z^{(1)}\right] \approx \left\{1 - \frac{\overline{h}}{h(Z^{(1)})}\right\} \mathbb{E}\left[N(t) \mid Z^{(1)}\right], \tag{4.10}$$

où

$$\overline{h} = \frac{\mathbb{E}\left\{\exp\left[f(Z^{(1)})\right] Y(t)\right\}}{\mathbb{E}[Y(t)]}$$

Cette dernière équation s'interprète naturellement : le nombre attendu de décès supplémentaires est approximativement égal à 1 moins le taux de risque que multiplie le nombre attendu d'événements.

Puisque $\widehat{M}$ et N sont connus, on peut inverser 4.10 afin d'obtenir

$$f(Z^{(1)}) - \overline{f} \approx -\log\left\{1 - \frac{sm(\widehat{M}, Z^{(1)})}{sm(N, Z^{(1)})}\right\}, \tag{4.11}$$

où

- $\overline{f} = \log(\overline{h})$,
- $sm(\widehat{M}, Z^{(1)})$ est une estimation lissée (*smoothed*) de $\mathbb{E}\left[\widehat{M}(t) \mid Z^{(1)}\right]$, qui peut être obtenue en traçant le graphe lissé de $\widehat{M}$ en fonction de X,
- $sm(N, Z^{(1)})$ est l'analogue, concernant N, de la quantité précédente.

Therneau(1990)[23] démontrent que l'équation 4.11) peut, pour $t = 1$, être remplacée par l'approximation suivante [1] :

$$\mathbb{E}\left[\widehat{M}(t) \mid Z^{(1)}\right] \approx c\left\{f(Z^{(1)}) - \overline{f}\right\} \tag{4.12}$$

où c est le nombre total d'événements divisé par le nombre total de sujets.

Ainsi, un graphe lissé des $\widehat{M}_i$ suivant une covariable fournira une approximation de la forme fonctionnelle correcte à placer dans l'exponentielle du modèle de Cox.

Enfin, un avantage de 4.12 par rapport à 4.11 réside en son interprétation : l'axe des ordonnées est à l'échelle directe des décès supplémentaires.

Therneau et al. [23] ont mené une expérience −limitée− de simulations, qui a montré que l'approximation 4.12 est acceptable lorsque $\beta\mathbb{E}(N) < 2$.

1. À condition que f n'ait pas de variation extrême et que la dépendance de $\mathbb{E}(N \mid X)$ par rapport à X soit faible, comme par exemple en présence d'un taux modéré de censure.

4.5.3 Concernant la proportionnalité des risques vis-à-vis d'une covariable

4.5.3.1 Méthodes graphiques

Supposons que nous désirions tester la proportionnalité du modèle vis-à-vis de la covariable $Z^{(1)}$, après ajustement sur les autres covariables. Notons $Z = (Z^{(1)}, Z^{(2)})^t$ le vecteur des covariables, où $Z^{(2)}$ est le vecteur des $(p-1)$ covariables restantes.

On suppose qu'il n'existe pas d'interaction entre $Z^{(1)}$ et les autres covariables.

Première méthode Supposons que $Z^{(1)}$ prenne K valeurs possibles. Si $Z^{(1)}$ est continue, nous stratifions les données suivant K strates, notées $G_1, \ldots, G_K$. Si $Z^{(1)}$ est discrète, elle prend les valeurs $1, 2, \ldots, K$.

Si nous utilisons un modèle de Cox stratifié, nous obtenons $\widehat{A}_{g0}(t), (g = 1, \ldots, K)$, qui est le risque de base cumulé pour la ge strate.

S'il y a proportionnalité des risques, le risque de base cumulé de chaque strate doit être un multiple des autres. Aussi, si nous traçons $\ln\left[\widehat{A}_{10}(t)\right], \ldots, \ln\left[\widehat{A}_{K0}(t)\right]$ en fonction de t, nous devons obtenir des courbes parallèles et l'écart entre ces courbes doit demeurer constant.

Nous pouvons aussi tracer $\ln\left[\widehat{A}_{g0}(t)\right], \ldots, \ln\left[\widehat{A}_{10}(t)\right]$ $(g = 2, \ldots, K)$ en fonction de t : nous devons alors obtenir des courbes globalement constantes.

Deuxième méthode Andersen (1982) propose, pour chaque t, de tracer $\widehat{A}_{g0}(t)(g = 2, \ldots, K)$ en fonction de $\widehat{A}_{10}(t)$. En cas de proportionnalité, ces courbes doivent être des droites passant par l'origine.

De plus, si $A_{g0}(t) = e^{\gamma g}$, alors la pente de ces droites devrait approximativement être une estimation de $e^{\gamma g}$.

Gill et Schumacher (1987) ont montré que si le graphe de $\widehat{A}_{g0}(t)$ en fonction de $\widehat{A}_{10}(t)$ est convexe (respectivement concave), alors le rapport $\alpha_{g0}(t)/\alpha_{10}(t)$ est une fonction croissante (resp. décroissante) de t.

Concernant ces deux premières méthodes, l'interprétation qui peut en être faite doit être considérée avec précaution, en raison des variances des courbes qui ne sont pas constantes au cours du temps.

Troisème méthode Arjas (1988) propose la méthode suivante.
Notons $T_{(1)} < T_{(2)} < \cdots < T_{(M)}$ ($M \leq n$, où n est le nombre de sujets) les temps de survenue de l'événement. Soit

$$\begin{aligned}\mathfrak{M}_i(k,\beta) &= N_i(T_{(k)}) - \int_0^t \frac{Y_i(s)\exp(\beta^t Z_i)}{\sum_{j=1}^n Y_j \exp(\beta^t Z_i)} d\left[\sum_{i=1}^n N_i(s)\right] \\ &= N_i(T_{(k)}) - \sum_{j\leq k} \frac{Y_i(T_{(j)})\exp(\beta^t Z_i)}{\sum_{l=1}^n Y_l(T_{(j)})\exp(\beta^t Z_l)}\end{aligned}$$

Si nous notons

$$H(k,\beta) = \sum_{i=1}^{n} \sum_{j \leq k} \frac{Y_i(T_{(j)}) \exp(\beta^t Z_i)}{\sum_{l=1}^{n} Y_l(T_{(j)}) \exp(\beta^t Z_l)}$$

alors

$$\begin{aligned} \widehat{\mathfrak{M}}(k,\beta) &= \sum_i \mathfrak{M}_i(k,\beta) \\ &= k - H(k,\beta) \end{aligned}$$

est une martingale.

Par suite, un graphe de $H(k,\beta)$ en fonction de k peut être comparé à la première bissectrice.

Quatrième méthode Cette méthode est basée sur les **résidus du score**. Pour la k_e variable concernant le ie sujet, la définition du résidu du score est, dans le contexte du modèle de Cox,

$$U_{ik}(\widehat{\beta}, \infty) = \int_0^{\infty} [Z_{ik} - E_k(\widehat{\beta}, s)] d\widehat{M}_i(s),$$

où $E_j(\beta, s)$ est la j^e composante du vecteur $E(.,.)$ défini en 4.7.

Le modèle de Cox avec les p covariables est ajusté. Quand toutes les covariables sont fixées au temps 0, le résidu du score à un temps de survenue de l'événement donné vaut

$$H_{ik} = \delta_i Y_i(t)[Z_{ik} - E_k(\widehat{\beta}, T_i)] - \sum_{t_{(i)} \leqslant t} [Z_{ik} - E_j(\widehat{\beta}, t_i)] Y_i(t_{(i)}) \exp(\widehat{\beta}^t)[\widehat{A}_0(t_{(i)}) - \widehat{A}(t_{t_{(i-1)}})], \tag{4.13}$$

où $0 = t_{(0)} < t_{(1)}) < \cdots < t_{(M)}$ sont, comme précédemment, les temps ordonnés de survenue de l'événement d'intérêt.

En utilisant les scores relatifs aux n individus, nous définissons un processus du score pour la ke covariable égal à

$$U_k(t) = \sum_{i=1}^{n} H_{ik}(t).$$

Le processus des scores est la première dérivée partielle de la fonction de vraisemblance partielle du modèle de Cox ajusté, qui utilise uniquement l'information disponible au temps t. Il est clair que $U_k(0) = 0$ et que $U(\infty) = 0$, puisque la valeur de β utilisée lors de la construction des résidus du score est la solution du vecteur d'équation $U_k(\infty) = 0$, $k = 1, \ldots, p$.

Si l'adéquation est correcte, alors le processus

$$W_k(t) = U_k(t) \times \sqrt{\mathbb{V}(\widehat{\beta}_k)}$$

converge vers un pont brownien fluctuant aux alentours de 0 (à condition que $\mathbb{C}ov(\widehat{\beta}_k, \widehat{\beta}_{k'} = 0$ pour $k \neq k'$). Par suite, un graphe de $W_k(t)$ en fonction du temps devrait ressembler à une marche aléatoire fluctuant autour de 0. Si les risques pour différents niveaux de la covariable ne sont pas proportionnels, alors les graphes doivent avoir un maximum qui est trop grand en valeur absolue, à un instant donné.

L'utilisation des résidus du score pour s'assurer de la proportionnalité des risques présente deux avantages par rapport aux autres approches :

1^0 les covariables continues sont traitées naturellement et n'ont pas besoin d'être discrétisées ;

2^0 un seul mod'ele de Cox doit être ajusté pour vérifier la proportionnalité des risques concernant toutes les covariables du modèle.

Cependant, du fait que la puissance du processus du score à détecter une non-proportionnalité des risques n'a pas été comparée à celle des graphiques d'Andersen ou d'Arjas, il est recommandé de mettre en pratique toutes les méthodes possibles.

4.5.3.2 Méthodes analytiques

La proportionnalité des risques concernant la covariable Z_j ne sera pas vérifiée si la statistique

$$\sup_t \sum_i \int_0^t \left[Z_{ij} - E_j(\widehat{\beta}, s) \right] d\widehat{M}_i(s)$$

dépasse une certaine valeur (rappelons que $E_j(\widehat{\beta}, s)$ est la j^e composante du vecteur défini en 4.7) .

Cette statistique devrait être très sensible aux alternatives pour lesquelles les covariables ont un comportement (de croissance ou de décroissance) monotone au cours du temps, et tout spécialement pour les alternatives telles que

$$\frac{h(t; Z = x)}{h(t; Z = y)}$$

soit strictement croissant en fonction de t pour tout $x < y$, ou bien strictement décroissant pour tout $x < y$.

Pour obtenir la distribution de cette statistique, on démontre que, sous les deux hypothèses données ci-après,

$$\sup_t \sqrt{\mathfrak{I}^{-1}(\widehat{\beta}, \infty)_{jj}} \sum_i \int_0^t \left[Z_{ij} - E_j(\widehat{\beta}, s) \right] d\widehat{M}_i(s)$$

suit asymptotiquement la distribution de

$$\sup_{0 \leqslant t \leqslant 1} W^0(t).$$

où W^0 est un pont brownien.

Les deux hypothèses mentionnées ci-dessus sont les suivantes :

1^0 la j^e composante du vecteur des covariables satisfait l'hypothèse de proportionnalité des risques ;

2^0 $V(t)_{jk} = 0$ pour tout t, où $V(.)$ est la covariance asymptotique de

$$\frac{1}{\sqrt{n}}\left(\sum_i \int_0^t \left[Z_{i1} - E_1(\widehat{\beta}, s)\right] d\widehat{M}_i(s), \ldots, \sum_i \int_0^t \left[Z_{ip} - E_p(\widehat{\beta}, s)\right] d\widehat{M}_i(s)\right)$$

Cette dernière condition de nullité nécessite que la covariable Z_j soit orthogonale aux autres covariables.
L'estimateur consistant $n^{-1}\mathfrak{I}^{-1}(\beta, t)$ de $V(t)$ est la somme, au long des temps de décès survenus dans l'intervalle $[0,t]$, des covariances de Z à chaque instant de décès.
Par exemple, $V(t)_{jk} \approx 0$ dans les études d'intervention pour lesquelles la je covariable représente le traitement administré après "randomisation" − aussi longtemps que de fortes interactions entre traitement et facteur n'existent pas.
Les cas où cette hypothèse n'est pas vérifiée demeurent à l'étude.

4.5.4 Justesse du modèle pour chaque sujet

L'usage graphique des résidus permet d'apprécier la pauvreté de la prédiction individuelle. La taille du résidu individuel $\widehat{M}_i$ indique la justesse du modèle, avec une grande valeur positive pour un sujet qui "meurt trop tôt" , et inversement, une grande valeur négative pour un sujet qui "vit trop longtemps".

Dans le modèle de Cox, les résidus martingales sont fortement étirés (*skewed*) et cet étirement − cette queue− déforme l'apparence du graphique standard des résidus. Notons que si la valeur maximale d'un tel résidu est finie et vaut 1, sa valeur minimale possible est -1.
Il est alors presque impossible de détecter les données abérrantes (*outliers*).
C'est pourquoi il est préférable d'utiliser les **résidus de déviance (*deviance residuals*)**.

Ces résidus sont définis à partir des résidus martingales par

$$d_i(t) = sgn\left(\widehat{M}_i(t)\right)\left\{-2\left[\widehat{M}_i(t) + N_i(t)\log\left(\frac{N_i(t) - \widehat{M}_i(t)}{N_i(t)}\right)\right]\right\}^{\frac{1}{2}}$$

Il s'agit d'une transformation empirique des résidus martingales. Dans cette expression, la racine carrée tend à diminuer les résidus martingales grandement négatifs, tandis que la transformation logarithmique accroît les résidus martingales qui sont proches de l'unité. Ainsi, la distribution des résidus de déviance est plus symétrique autour de zéro que celle des résidus martingales.

Dans le cas du modèle de Cox, les résidus de déviance s'écrivent

$$d_i(T_i) = sgn\left(\widehat{M}_i(T_i)\right)\left\{-2\left[\widehat{M}_i(T_i) + \delta_i \log\left(\delta_i - \widehat{M}_i(T_i)\right)\right]\right\}^{\frac{1}{2}}$$

Le graphe des résidus de déviance en fonction des quantités

$$\sum_{k=1}^{p} \hat{\beta} Z_{ik}$$

doit, lorsque la censure est modérée, ressembler à un échantillon de bruit distribué suivant une loi normale. En cas de censure sévère, une grande quantité de points proches de 0 va déformer l'approximation normale.

Dans tous les cas, les valeurs aberrantes potentielles auront des résidus de déviance dont les valeurs absolues seront trop importantes.

4.5.5 Concernant les "observations influentes"

4.5.5.1 Méthodes graphique

Cette méthode consiste en la comparaison de l'estimation $\hat{\beta}$ −obtenue en estimant β à partir de toutes les observations − et de l'estimation $\hat{\beta}_{(i)}$ − obtenue en excluant des observations celle relative au i^e sujet.

Reprenons la définition des $H_{ik}(t)$ (4.13), et notons $H_{ik} = H_{ik}(\infty)$. Nous avons maintenant

$$H_{ik} = \delta_i \left[Z_{ik} - E_k(\hat{\beta}, T_i) \right] - \sum_{t_b \leqslant T_i} \left[Z_{ik} - E_j(\hat{\beta}, t_b) \right] \exp(\hat{\beta}^t) \left[\hat{A}_0(t_b) - \hat{A}_0(t_{b-1}) \right].$$

Le premier terme

$$\begin{aligned} R_{ik} &= \delta_i - \left[Z_{ik} - E_k(\hat{\beta}, T_i) \right] \\ &= \delta_i \left[Z_{ik} - \frac{\sum_{j=1}^{n} Y_j(t) Z_{jk} \exp(\hat{\beta} Z_j)}{\sum_{j=1}^{n} Y_j(t) \exp(\hat{\beta} Z_j)} \right] \end{aligned}$$

est le **résidu partiel de Schoenfeld** (Schoenfeld, 1982)[21] : il s'agit de la différence entre la valeur observée de la covariable Z_{ik} à l'instant de survenue de l'événement, et la valeur obtenue par le modèle à ce même instant.
Il est possible de montrer que $\hat{\beta} - \hat{\beta}_{(i)}$ est approximativement égal à

$$\mathfrak{I}(\hat{\beta})(H_{i1}, \ldots, H_{ip})^t,$$

où $\mathfrak{I}^{-1}(\hat{\beta})$ est la matrice d'information de Fisher observée.

Le graphe −pour chaque covariable− de ces résidus de Schoenfeld en fonction soit des temps de survenue de l'événement, soit de la covariable Z_{ik}, est utilisé pour ètablir l'influence de la i^e observation sur la k^e covariable (les R_{ik} en fonction des T_i doivent être centrés autour de 0).

4.5.5.2 Méthode analytique

L'influence d'une observation sur le modèle dépend à la fois du résidu obtenu après ajustement et de la valeur extrême de la covariable, soit grossièrement $Z_i - E(\beta,t)$ que multiplie le résidu. $E(\beta,t)$ est une fonction du temps : c'est la moyenne sur l'ensemble des individus à risque au temps t. Ceci suggère d'utiliser une valeur "moyenne au cours du temps" de $Z_{ij} - E_j(\beta,t)$, et finalement l'on parvient au résidu du score

$$\int_0^\infty (Z_{ij} - E_j(\hat{\beta}_j,s))d\widehat{M}_i(s)$$

comme outil de mesure de l'influence d'une observation.

Une manière de formaliser ce résultat est d'ajouter des poids aux observations individuelles afin de donner une vraisemblance partielle et un vecteur de score pondérès. Ainsi [2],

$$\begin{aligned}\frac{\partial\hat{\beta}}{\partial\omega_i} &= \left(\frac{\partial\hat{\beta}}{\partial U}\right)\left(\frac{\partial U}{\partial\omega_i}\right)\\ &= \mathfrak{I}(\hat{\beta})^{-1}\frac{\partial U}{\partial\omega_i},\end{aligned}$$

calculée au point $\omega = 1$, est l'estimateur *jackknife* infinitésimal de l'influence de la i^e observation sur β . Une manipulation algébrique révèle que le second terme de l'équation ci-dessus est exactement le résidu du score, si bien que le vecteur d'influence du i^e sujet est

$$-\mathfrak{I}(\hat{\beta})^{-1}\left(\int_0^\infty [Z_{i1} - E_1(\hat{\beta},s)]d\widehat{M}_i(s),\ldots,\int_0^\infty [Z_{ip} - E_p(\hat{\beta},s)]d\widehat{M}_i(s)\right)^t$$

Cette méthode sous-estime le *jackknife* réel, spécialement pour des valeurs extrêmes de Z, puisque $\mathfrak{I}$ change également quand l'observation est modifiée.

Une autre méthode consiste à calculer le "premier pas de l'actualisation" (*1-step update*) de $\hat{\beta}$ quand une seule covariable Z_{p+1} est ajoutée, avec Z_{p+1} valant 1 pour le sujet i et 0 pour tous les autres.
Ici, le changement au premier pas, au point $(\hat{\beta},0)$ vaut

$$\Delta\hat{\beta}_{(i)} = \frac{-\mathfrak{I}(\hat{\beta})^{-1}\gamma_i}{\eta_i - \gamma_i'\mathfrak{I}(\hat{\beta})^{-1}\gamma_i}\widehat{M}_i,$$

où

$$\gamma_{ij} = \int_0^\infty Y_i(s)[Z_{ij} - E_j(\hat{\beta},s)]\exp(\hat{\beta}^t Z_i(s))d\hat{\Lambda}_0(s)$$

et

$$\eta_i = \int_0^\infty Y_i(s)[1 - E_{p+1}(\hat{\beta},s)]\exp(\hat{\beta}^t Z_i(s))d\hat{\Lambda}_0(s)$$

Cette expression est très similaire à celle des estimateurs *jackknife* du modèle linéaire, avec $\widehat{M}_i$ considéré comme résidu.

Remarque 4.3 Les tests usuels d'adéquation −Kolmogorov-Smirnov, von Mises − peuvent être adaptés au cas où les covariables dépendent du temps (Marzec et Marzec, 1997).

2. En reprenant la notation de 4.3.2.1

4.6 Considération des ex-aequo

L'expression de la vraisemblance partielle de Cox (cf. 4.4) ne vaut que sous l'hypothèse de temps de survenue de l'événement distincts, c'est-à-dire lorsque $\Delta N_i(t) = N_i(t) - N_i(t-)$ ne peut valoir que 0 ou 1, quel que soit i.

Cependant, des adaptations de cette vraisemblance partielle ont été conçues, afin de traiter des données de survie présentant des ex-aequo ; nous les donnons ci-dessous.

4.6.1 Vraisemblance partielle de Breslow

Elle vaut

$$L(\beta) = \prod_t \prod_i \frac{Y_i(t)\exp(\beta^t Z_i)}{\left[S^{(0)}(\beta,t)\right]^{\Delta N_i(t)}}.$$

4.6.2 Vraisemblance partielle d'Efron

Elle est de la forme

$$L(\beta) = \prod_t \prod_i \left\{ \frac{Y_i(t)\exp(\beta^t Z_i)}{\prod_{k=1}^{\Delta N_i(t)} \left[\sum_j Y_j(t)\exp(\beta^t Z_j) - \frac{k-1}{\Delta N_i(t)} \sum_l Y_l(t)\delta_l \exp(\beta^t Z_l\right]^{\Delta N_i(t)}} \right\}.$$

4.6.3 Vraisemblance partielle exacte

Son expression est

$$L(\beta) = \prod_t \prod_i \left[\int_0^\tau \prod_{j=1}^{\Delta N_i(t)} \left(1 - \exp\left\{ - \left[\frac{\exp(\beta^t Z_j}{\sum_l Y_l(t)\delta_l \exp(\beta^t Z_l} \right] t \right) \right\} \exp(-t) dt \right].$$

4.7 Extensions du modèle

Le modèle de Cox peut être étendu à plusieurs cas, notamment ceux concernant une stratification des covariables, ou bien encore une dépendance de ces mêmes covariables vis-à-vis du temps.

Il est également possible, dans le cas de données emboîtées (où T_{ij} est le temps de survie de l'individu j appartenant au groupe i), de diversifier la fonction de risque de base : $h_0(t)$ devient $h_{i0}(t)$, fonction de risque de base propre au groupe i.

Il est enfin possible d'introduire une corrélation entre les données de survie . Notons que le modèle de Cox − tout comme un autre modèle, dit à *temps accélérés* − est un cas particulier d'un modèle ayant vu le jour en 1987 : le **modèle étendu de régression du risque instantané** (***extended hazard regression***) (Etezadi-Amoli et Ciampi). Ce modèle spécifie que le risque instantané s'écrit

$$h(t) = g_1(\beta^t Z) \times h_0[g_2(\gamma^t Z)t],$$

où $g_1(x)$ et $g_2(x)$ sont des fonctions positives égales à 1 en 0, $h_0(t)$ est le risque de base, et β et γ sont les paramètres de la régression.

4.7.1 Covariables dépendantes du temps

Le modèle de Cox permet de prendre en compte des covariables dépendantes du temps (traitement, marqueur biologique,...). Il faut néanmoins que Z(t) soit prÃ©dictible, c'est-è-dire connue au temps t : Le traitement statistique est identique néanmoins, on peut faire les remarques suivantes.

- Il est nécessaire de connaître la valeur des covariables pour chaque temps d'événement.
  En effet, on a $L_{cox}(\beta) = \prod_{i=1}^{D} = \frac{exp\left(\beta' Z_{(i)}(T_i)\right)}{\sum_{j \in R(T_i)} exp\left(\beta' Z_j(T_i)\right)}$.Ceci peut poser quelques problèmes dans le cas de marqueurs biologiques.
- L'interprétation devient difficile car le risque est spécifique à chaque histoire des covariables.
- L'hypothèse de hasard proportionnel est conservée. En effet, les fonctions de risque pour les différentes modalités d'une covariable restent proportionnelles et leurs rapports sont indépendants du temps. L'effet de la covariable ne varie pas au cours du temps, c'est la variable qui varie.
- L'utilisation de certaines covariables dépendantes du temps permet de tester l'hypothèse de risques proportionnels.

4.7.2 Modèle de Cox stratifié

Dans le cas où une variable qualitative ne vérifie pas l'hypothèse de hasards proportionnels, on peut considérer un modèle de Cox stratifié. Prenons l'exemple, d'une variable binaire Y codée 0 et 1, par exemple, le sexe (0 pour les hommes et 1 pour les femmes).
Dans ce modèle, le risque de base est différent dans les deux strates mais les covariables Z agissent de la même manière sur les deux fonctions de hasard, c'est-à-dire,

$$h(t \mid Z, Y = 0) = h_0(t)\exp(\beta' Z),$$

$$h(t \mid Z, Y = 1) = h_1(t)\exp(\beta' Z),$$

L'effet des covariables est le même dans chaque strate. Les estimations obtenues par la méthode de la vraisemblance partielle sont applicables pour obtenir les paramètres (h_0, h_1 et β) du modèle. La vraisemblance partielle est calculée dans chacune des strates ; la vraisemblance totale est le produit des vraisemblances de chaque strate.
Le modèle de Cox stratifié fait l'hypothèse que les covariables Z agissent de la même manière dans chaque strate. Cette hypothèse peut être testée en utilisant le test du rapport de vraisemblance :

$$\chi^2_{LRT} = 2\left[\sum_{j=1}^{s} \log\left(L_{Cox}(\hat{\beta}_j)\right) - \log\left(L_{Cox}(\hat{\beta})\right)\right]$$

où s est le nombre de strates, p le nombre de covariables (dimension de $\hat{\beta}$),$\sum_{j=1}^{s} \log\left(L_{Cox}(\hat{\beta}_j)\right)$ est la log-vraisemblance en considérant un β différent dans chaque strate et $\log\left(L_{Cox}(\hat{\beta})\right)$ est la vraisemblance en considérant le même β dans chacune des strates.

4.7.3 Modèles de fragilité (frailty)

Le modèle de Cox (et les méthodes traditionnelles) suppose que la population est homogène (malgré la prise en compte de covariables). Néanmoins, cette hypothèse n'est pas toujours réaliste, notamment que des covariables importantes ne sont pas observables ou inconnues. Par exemple, cela peut être des facteurs environnementaux ou génétiques. Les modèles fragilité permettent de prendre en compte l'hétérogénéité des observations.

Considérons une nouvelle covariable non observée Z_0. On suppose comme dans le modèle de Cox, que l'effet des covariables se résume à une quantité réelle $\exp(\beta_0 Z_0$, alors la fonction de risque est

$$h(t \mid Z, Z_0) = h_0(t) e^{\beta_0 Z_0} e^{\beta' Z}.$$

En notant $\omega = e^{\beta_0 Z_0}$, la variable aléatoire réelle positive (appelée fragilité), la fonction de risque devient

$$h(t \mid Z, \omega) = h_0(t) \omega e^{\beta' Z},$$

et la fonction de survie conditionnelle est

$$S(t \mid Z, \omega) = \exp\left(- \int_0^t h_0(s) \omega e^{\beta' Z} ds \right) = \exp\left(-\omega e^{\beta' Z} H_0(t) \right).$$

Comme ω est une variable aléatoire, on s'intéresse à la fonction de survie moyennée sur ω. Cette quantité correspond à la fonction de survie marginale pour un individu quelconque.

$$S(t \mid Z) = \int_0^t \exp\left(-v e^{\beta' Z} H_0(t) \right) f_\omega(v) dv = \mathcal{L}_\omega \left(e^{\beta' Z} H_0(t) \right)$$

où f_ω représente la densité de la v.a. ω et $\mathcal{L}_\omega(s) = \mathbb{E}(e^{-\omega s})$ est la transformée de Laplace de la distribution de la fragilité.

Le plus souvent, les modèles de fragilité sont utilisés pour prendre en compte une dépendance entre les temps d'événements de certains individus. En effet, les individus d'un même sous-groupe d'une population peuvent être liés si tous les individus de ce groupe ont des caractéristiques communes non observées. Par exemple, des individus d'une même famille, d'une même région ou d'un même hôpital. Le terme fragilité est alors commun à chaque individu du groupe (permet de créer la dépendance) mais différent d'un groupe à l'autre (hétérogénéité d'un groupe). on parle de modèle à fragilités partagées (shared frailty model).
Considérons $T_{ij} = \min(X_{ij}, C_{ij})$ où j représente l'indice du $j^{ème}$ individu du groupe $i (i = 1, ..., G)$. Le risque pour l'individu j du groupe i est

$$h_{ij}(t \mid Z_{ij}, \omega_i) = h_0(t) \omega_i e^{\beta' Z_{ij}}, \tag{4.14}$$

où ω_i est la fragilité du groupe i. Les ω_i sont $i.i.d.$ et en général, on suppose que $\mathbb{E}(\omega_i) = 1$ et $\mathbb{V}(\omega_i) = \theta$ pour des questions d'identifiabilité du modèle (dans ce cas si $\omega_i > 1$, le risque du groupe i sera supérieur en moyenne au risque de base et inversement si $\omega_i < 1$).
Le paramètre θ permet alors de mesurer l'hétérogénéité entre les groupe (une variance importante entraîne une grande variabilité entre groupes).

- Dans ce modèle, les observations sont indépendantes conditionnellement aux ω_i.
- La loi Gamma est souvent utilisée comme loi des effets aléatoires car elle a de bonnes propriétés mathématiques. Elle fournit de bons résultats en pratique les $r^{èmes}$ dérivées de la transformée de Laplace ont une écriture simple.
- D'autres distributions sont possible : loi inverse gaussienne, loi positive stable...
- On utilise souvent l'algorithme EM pour estimer les paramètres du modèle. On peut également passer par la transformée de Laplace dans le cas de la loi Gamma.
- Le modèle 4.14 implique que les risques sont proportionnels conditionnellement aux valeurs de fragilité. Par conséquent, l'interprétation de β est conditionnelle à la fragilité. Par exemple, si $Z_{ij} = 0$ ou 1, cela signifie que e^{β} représente le risque entre un sujet codé 1 et un sujet codé 0 au sein d'un même groupe.
- Plusieurs généralisations sont possibles :
  - modèle de fragilité stratifié : des risques de base différents dans chaque strate, par exemple, pour différencier les hommes et les femmes au sein d'une même famille ;
  - modèle avec une fragilité qui suit une loi multivariée (par exemple pour prendre en compte le fait que dans une famille, les frères et les soeurs sont plus proche entre eux que les cousins) ;
  - modèle avec deux fragilités pour prendre en compte une dépendance causée par deux sortes de raisons (facteurs génétiques et facteurs environnementaux) ;
  - modèle avec une fragilité dépendante du temps.

Remarque 4.4 *Les modèles de fragilités peuvent être utilisés pour prendre en compte l'hétérogénéité entre les individus d'une population. Dans ce cas, il y'a une valeur ω_i par individu $(i = 1, \ldots, n)$,*

$$h_i(t \mid Z_i\omega_i) = h_0(t)\omega_i e^{\beta' Z_i}.$$

Le paramètre $\mathbb{V}(\omega_i) = \theta$ mesure l'hétérogénéité entre les individus.

Chapitre 5

Régression Logistique

Sommaire

5.1 Introduction

La régression ordinaire permet d'analyser une variable réponse quantitative en fonction d'une ou plusieurs variables explicatives. Souvent, c'est un résultat binaire (ou dichotomique) d'une expérience ou d'une observation que l'on souhaite mettre en relation avec des variables explicatives ; par exemple :

- des patients peuvent survivre ou décéder ; les différentes thérapies et les facteurs de risque peuvent être considérés comme des variables qui contribuent à expliquer la survie ou le décès ;
- des personnes peuvent être atteintes par une maladie. On souhaite étudier la relation entre les chances d'être atteint et certains facteurs explicatifs ou facteurs de risque (par exemple, âge, fumée, sexe) ;
- des personnes peuvent avoir ou ne pas avoir un emploi selon leur âge, sexe, type de formation ;
- un appareil peut fonctionner ou ne pas fonctionner ; cet état peut être mis en relation avec son âge, les conditions de l'environnement, etc.

La régression logistique permet d'étudier la relation entre une variable réponse binaire et plusieurs variables explicatives.

Historiquement, l'étude des modèles décrivant les modalités prises par une ou plusieurs variables qualitatives date de la période des années 1930-1950. Les travaux les plus marquants de cette époque sont ceux de Bliss (1935) et de Berkson (1944). Ces travaux traitent les modèles dichotomiques simples (modèles logit et probit). Les premières applications ont été alors menées dans le domaine de la biologie. Ainsi, ce n'est que récemment, que les modèles multinomiaux ont été utilisés. Nous pouvons citer, à titre d'exemple, les travaux de MacFadden et al. (1978) qui ont développé le modèle logit multinomial. Mais, ce modèle présente une principale limite connue sous le nom de l'hypothèse d'indépendance des alternatives non pertinentes. En effet, un autre modèle plus flexible et acceptant n'importe quelle structure d'erreurs a été développée par Daganzo (1979). C'est le modèle probit multinomial.

En tant que procédure non paramétrique, la régression logistique présente l'avantage de ne pas exiger de contraintes quant à la normalité des distributions des variables. La régression logistique est moins une méthode d'inférence statistique qu'une méthode de classification. En effet, l'équation étudiée traduit la probabilité d'appartenance d'un sujet à une catégorie ou un groupe. Tous les modèles logistiques permettent d'analyser des situations réelles. Ils peuvent nous donner des résultats très précieux. Ces modèles sont utilisés surtout pour expliquer certaines maladies qui sont jugées très compliquées.

La première méthode, appelée ***régression logistique binaire*** (binary logistic regression), correspond au cas où la variable Y comporte uniquement deux classes, les individus étant décrits par la présence ou l'absence d'un caractère donné. Par exemple, des individus (parcelles, plantes, animaux, etc.) peuvent être attaqués ou non par un parasite, être fertiles ou non, être porteurs ou non d'une tare, etc.

La deuxième méthode, appelée ***régression logistique polychotomique nominale*** (polytomous nominal logistic regression), permet de traiter les cas où la variable à expliquer possède plus de deux classes si celles-ci ne peuvent pas être ordonnées ou si on ne souhaite pas tenir compte de l'ordre dans le cas où elles seraient ordonnées. Une telle situation se présente par exemple si des individus sont caractérisés par l'appartenance à une espèce donnée, par une couleur ou par le choix d'une réponse à une question posée parmi trois propositions telles que « oui », « non », « ne sait pas ».

Enfin, la troisième méthode, appelée ***régression logistique polychotomique ordinale*** (polytomous ordinal regression), concerne les situations où la variable y présente plus de deux modalités qui peuvent être ordonnées et dont on souhaite tenir compte de l'ordre. Un exemple typique est la description de l'intensité de l'attaque d'individus par un parasite, cette description étant réalisée par exemple sur la base d'une échelle à quatre niveaux notés A, B, C et D, le niveau A représentant l'absence d'attaque, le niveau B une attaque faible, le niveau C une attaque modérée et le niveau D une attaque forte.

Dans ce chapitre, après avoir justifié la nécessité de modèles particuliers, nous présenterons la régression logistique simple (avec comparaison de lien logit et probit), puis la régression logistique multiple. Nous terminerons par un résumé des tests de validité générale du modèle de régression logistique et quelques recommandations concernant la construction de tels modèles.

Notations

- $X = (X_1, X_2, .., X_p)$:variables aléatoires explicatives de dimension p, $x = (x_1, x_2, , x_p)$ une réalisation de X.
- Y une variable aléatoire à expliquer, à K modalités.
- $(X_1, Y_1), .., (X_n, Y_n)$: un n-échantillon aléatoire (iid et de même loi que le couple (X, Y)) tel que $X_1 = (X_{i1}, X_{i2}, .., X_{ip})$.
- $x = (x_1, y_1), ..., (x_n, y_n)$ une réalisation de $(X_1, Y_1), .., (X_n, Y_n)$.

5.1.1 Rappel sur le modèle linéaire

Nous cherchons à expliquer une variable Y par p variables $X = (X_1, X_2, .., X_p)'$. Pour se faire, on dispose de n réalisations $(x_1, y_1), ..., (x_n, y_n)$ du couple (X, Y). Le but est de modéliser la dépendance de la variable réponse Y sur les variables explicatives$X_1, X_2, .., X_p$. Plusieurs raisons peuvent motiver cette modélisation :

- la description : on veut un modèle qui permette de décrire la relation entre Y et X ;
- l'évaluation des contributions relatives de chaque prédicteur pour expliquer Y ;
- la prédiction : prévoir la valeur de Y pour des nouvelles variables explicatives. Le modèle linéaire classique s'écrit :

$$Y = X'\beta + \varepsilon = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon, \tag{5.1}$$

Avec $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. On distingue alors deux cas :
- Les variables X_i sont déterministes (non-aléatoires).
- Les variables X_i sont aléatoires.

5.2 Pourquoi les modèles particuliers ?

Considérons par exemple la population des ménages et intéressons nous à la variable Y prenant deux valeurs :
- $Y = 1$ si le ménage est propriétaire de sa résidence principale,
- $Y = 0$ si le ménage n'est pas propriétaire de sa résidence principale.

La variable de réponse à expliquer $Y_i \mid x_i$ suit une loi de Bernoulli de paramètre π_i :

$$\pi_i = P(Y_i = 1 \mid x_i)$$

Où x_i est le vecteur ligne des valeurs observées pour les variables explicatives. L'objectif est de construire un modèle permettant de reconstituer $E(Y_i \mid x_i)$, c'est-à-dire π_i en fonction des variables explicatives.
La nécessité de modèles particuliers se justifie par plusieurs considérations :
-l'utilisation d'un modèle de régression linéaire classique n'est pas adéquate,
-L'introduction de variables latentes amène naturellement à la modélisation proposée dans le cadre des modèles linéaires généralisés.

5.2.1 Le modèle de régression linéaire usuel est inadapté

Dans le cadre de la régression linéaire classique :

$$Y_i = x_i'\beta + \varepsilon_i \tag{5.2}$$

Y_i est une variable aléatoire quantitative,
$x_i'\beta$, prédicteur liéaire, c'est un élément déterminé,
ε_i est une variable aléatoire que l'on suppose telle que :

$$E(\varepsilon_i \mid x_i) = 0 \tag{5.3}$$

$$V(\varepsilon_i \mid x_i) = \sigma^2 \tag{5.4}$$

Si de plus les variables ε_i sont supposées gausiennes, l'estimateur des moindres carrés ordinaire :

$$\widehat{\beta} = (X'X)^{-1}X'y$$

est l'estimateur du maximum de vraisemblance.
Si l'on utilise le modèle de régression usuelle pour une variable dichotomique, l'equation (5.2) nous indique que le résidu serait distribué selon une loi discrète prenant deux valeurs :

$$\varepsilon_i = 1 - x_i\beta, \quad \text{avec la probabilité} \quad \pi_i$$

$\varepsilon_i = -x_i\beta$, avec la probabilité $1-\pi_i$

Ce qui est trop éloigné des hypothèses usuelles de continuité et de normalité des résidus. Par ailleurs, l'estimateur $\widehat{\beta}$ n'est plus efficace.
L'équation (5.2) implique que $E(Y_i) = x_i\beta$. Or Y_i suit de Bernoulli de paramètre π_i et d'espérance $E(Y_i) = \pi_i$. Il en découle que $\pi_i = x_i\beta$ or rien n'implique que $x_i\beta$ sera compris entre 0 et 1.

5.2.2 Variables latentes

Les méthodes proposées partent du principe que le phénomène étudié caractérisé par l'observation d'une variable dichotomique est la manifestation visible d'une variable latente Z inobservable qui, elle, est continue.

Prenons l'exemple de la possession d'un bien durable par un ménage.La variable latente peut être « l'intensité du désir » de posséder le bien. Tant que cette intensité reste inférieure à un certain seuil, on observe $Y_i = 0$ (le ménage i ne possède pas le bien), quand elle le dépasse, on observe $Y_i = 1$(le ménage i possède le bien).

On peut aussi formuler le problème en terme de fonction d'utilité : pour le ménage i de caractéristiques x_i(âge de la personne de référence, sexe, revenu,CSP,...), la possession du bien procure un niveau d'utilité $U(1,x_i)$, alors que la non possession procure un niveau $U(0,x_i)$.

On a alors :

$$Y_i = 1 \Leftrightarrow U(1,x_i) > U(0,x_i),$$
$$Y_i = 0 \Leftrightarrow U(0,x_i) > U(1,x_i).$$

Le ménage choissisit la situation qui lui procure le plus haut niveau d'utilité.

On se ramène au cas de la variable latente en posant :

$$Z_i = U(1,x_i) - U(0,x_i).$$

On en déduit alors :

$$Y_i = 1 \Leftrightarrow Z_i > 0,$$
$$Y_i = 0 \Leftrightarrow Z_i < 0.$$

Il y a la possession du bien, lorsque la varibale latente Z_i dépasse le seuil 0.

5.2.3 Modèle théorique prenant en compte la présence d'une variable latente

Disposant d'une variable dichotomique Y à expliquer, notons Z la variable latente sous-jacente au phénomène étudié, le modèle postule une relation du type :

$$Z_i = x_i\beta + \varepsilon_i,$$

Où x_i est le vecteur contenant l'ensemble des valeurs des variables explicatives.

La probabilité que la variable Y_i prenne la modalité 1 est alors $\pi_i = P(Y_i = 1 \mid x_i) = P(Z_i > 0)$. En tenant compte de la modélisation de Z_i, cette probabilité s'écrit $P(x_i > -\varepsilon_i) = F(x_i\beta)$ si l'on note F la fonction de répartition de ε_i. Le choix du modèle porte donc sur le choix de cette fonction de répartiton. Deux fonctions de répartition sont couramment utilisées :

– la fonction de répartition de la loi logistique,
– la fonction de répartition de la loi normale.

Dans le cadre des modèles linéaires généralisés, comme le montre l'expression :

$$\pi_i = P(Y_i = 1 \mid x_i) = E(Y_i \mid x_i) = F(x_i\beta),$$

Ces fonctions représentent l'inverse de la fonction de lien et conduisent respectivement à la définition de modèles logit et de modèle probit.

5.2.4 Justification concernant le choix de la fonction logistique

La fonction logistique est définie par :

$$F(x) = \frac{e^x}{1+e^x}$$

Cette fonction est bien adaptée à la modélisation de probabilités, car elle prend ses valeurs entre 0 et 1 selon une courbe en S (Figure 5.1). son utilisation est par exemple indiquée lors de la modélisation du risque individuel de développer une maladie dans les études épidémiologiques. En effet, en considérant que la variable X représente un indice résultant de la combinaison de plusieurs facteurs de risque, on peut interpréter $F(x)$ comme le risque d'être atteint de cette maladie. Dans ce contexte le risque minimal pour de faibles valeurs de x, il augmente pour les valeurs intermédiaire de x et apparaît proche de 1 pour des valeurs plus élevées de x.

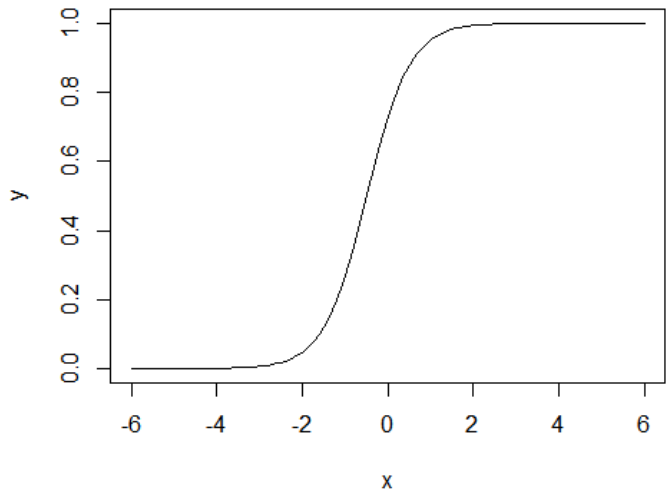

FIGURE 5.1 – Fonction logistique

5.2.5 Les modèles logit et probit

5.2.5.1 Modèle logit :

Le modèle logit est celui pour le quel

$$F(x) = \frac{e^x}{1+e^x}$$

est la fonction de répartition de la loi logistique de moyenne 0, et de variance $\frac{x^2}{3}$.

La fonction inverse est définie par :

$$F^{-1}(t) = \log\left(\frac{t}{1-t}\right)$$

5.2.5.2 Modèle probit :

Le modèle probit est celui pour le quel

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

est la fonction de répartition de la loi normale centrée réduite. $F^{-1} = \Phi^{-1}$ est la fonction probit.

5.2.6 Comparaison des deux modèles

L (fonction de répartition de la loi logistique) et Φ (fonction de répartition de la loi normale) sont toutes les deux symétriques par rapport à 0, et comprises entre 0 et 1(ce qui convient pour représenter une probabilité).

La loi logistique de fonction de répartition L a pour moyenne 0 et de variance $\frac{\pi^2}{3}$; il est donc naturel de comparer à $\Phi(x)$, fonction de répartition de $N(0,1)$, la fonction $L(x)$:

$$L(x) = \frac{1}{1+\exp(-\pi x/\sqrt{3})}$$

Dans la plupart des cas pratiques, on peut choisir indifféremment l'un ou l'autre modèle. Le modèle LOGIT a l'avantage d'une plus grande simplicité numérique. Le modèle PROBIT est en revanche plus proche du modèle habituel de régression par les moindres carrés.

5.3 Le modèle de régression logistique dichotomique

On se place dans le cas où Y prend deux modalités (0 ou 1, présence ou absence d'une maladie, panne ou non d'un composant électronique, bon ou mauvais client...). Nous représenterons ces deux modalités par 0 et 1 dans la suite. La modalité 1 est généralement utilisée pour le caractère que l'on cherche à étudier (achat d'un produit, présence d'une maladie, panne...). Le modèle de régression vus précédemment ne s'applique plus puisque le régresseur linéaire habituel $X\beta$ ne prend pas des valeurs simplement binaire.

5.4 Le modèle

L'idée est alors de ne plus modéliser Y, mais les probabilités d'avoir $Y = 0$ et $Y = 1$ conditionnellement à la connaissance des variables explicatives $X = x$:

$$\pi(x) = P(Y = 1 \mid X = x) \qquad et \qquad 1 - \pi(x) = P(Y = 0 \mid X = x)$$

Même si π n'est plus binaire, elle est toujours bornée dans l'intervalle $[0,1]$, ce qui ne convient toujours pas à un régresseur linéaire $X\beta$ qui prendra à priori des valeurs sur tout $\mathbb{R}$. La régression logistique consiste donc à modéliser une certaine transformation de π, appelée transformation logit, par une fonction linéaire des variables explicatives :

$$logit\left(\pi(x)\right) = \ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{j=1}^{p}\beta_j X_j \tag{5.5}$$

Ce modèle s'écrit également :

$$\pi(x) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j X_j\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j X_j\right)} \tag{5.6}$$

Dans la suite, nous noterons parfois $\pi(x;\beta)$ pour signifier que la probabilité $\pi(x)$ est paramétrée par β, et de même $P(Y=1|X=x;\beta)$.

5.5 Odds et odds-ratio

Le succès de la régression logistique, très utilisée en entreprise (finance, assurance, médecine, marketing ...), est en partie dû aux capacités d'interprétabilité du modèle. On définit par ***odds*** le rapport :

$$odds(x) = \frac{\pi(x)}{1-\pi(x)}$$

qui représente combien de fois on a plus de chance d'avoir $Y=1$ au lieu d'avoir $Y=0$ lorsque $X=x$.

On définit de même les ***odds-ratio*** par le rapport :

$$odds-ratio(x_i,y_j) = \frac{odds(x_i)}{odds(y_j)}$$

Qui représente combien de fois on a plus de chance d'avoir $Y=1$ au lieu d'avoir $Y=0$ lorsque $X=x_i$ au lieu de $X=x_j$.

Remarque 5.1 Bien que l'on ait défini les odds et odds-ratio pour une variable explicative X multidimensionnelle, on ne fait généralement varier qu'une seule dimension entre les deux valeurs x_i et x_j , et on définit donc autant d'odds et odds-ratio qu'il y a de dimensions.

Exemple 5.1 Exemple On considère comme variable à prédire Y la présence ou l'absence d'un cancer des poumons, et comme variable explicative (qualitative) le fait d'être fumeur ou non fumeur. Les données sont fictives bien que pas si éloignées que cela de la réalité :

-La probabilité d'avoir un cancer du poumon chez un fumeur est $P(Y=1|X=fumeur)=0.01$, d'où $P(Y=0|X=fumeur)=0.99$. On a alors $odds(X=fumeur)=1/99$. On dit que l'on a une chance sur 99 d'avoir un cancer des poumons lorsque l'on est fumeur.

5.6 Estimation des paramètres

5.6.1 Estimation des β_j

Les paramètres à estimer sont $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ Si on dispose d'un échantillon (y_i, x_i) $i = 1, .., n$,où $x_i = (x_i, ..., x_{ip})$, telle que les y_i soient indépendants conditionnellement aux x_i, on peut estimer β par maximum de vraisemblance.

Les probabilités de Y étant exprimées conditionnellement aux variables explicatives X, nous maximisons la vraisemblance conditionnelle :

$$L(\beta) = \prod_{i=1}^{n} P(Y = y_i \mid X = x_i)$$

Or, en utilisant la notation $\widetilde{x_t} = (1 x_i)'$, on a :

$$\begin{aligned} P(Y = y_i \mid X = x_i) &= \begin{cases} \dfrac{\exp \beta \widetilde{x_t}}{1 + \exp \beta \widetilde{x_t}} & si \quad y_i = 1 \\ 1 - \dfrac{\exp \beta \widetilde{x_t}}{1 + \exp \beta \widetilde{x_t}} & si \quad y_i = 0 \end{cases} \\ &= \left(\frac{\exp \beta \widetilde{x_t}}{1 + \exp \beta \widetilde{x_t}} \right)^{y_i} \left(1 - \frac{\exp \beta \widetilde{x_t}}{1 + \exp \beta \widetilde{x_t}} \right)^{1 - y_i} \end{aligned}$$

d'où la log-vraisemblance

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} P(Y = y_i \mid X = x_i) \\ &= \sum_{i=1}^{n} y_i \beta' \widetilde{x_t} - \ln(1 + \exp \beta' \widetilde{x_t}). \end{aligned}$$

La maximisation de cette vraisemblance se fait en dérivant par rapport au vecteur β. On obtient

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{n} y_i \widetilde{x_t} - \widetilde{x_t} \frac{\exp \beta \widetilde{x_t}}{1 + \exp \beta \widetilde{x_t}} \\ &= \sum_{i=1}^{n} \widetilde{x_t} (y_i - \pi(x_i)) \end{aligned}$$

qui n'est pas une équation linéaire en β. Sa résolution peut être réalisée numériquement par un algorithme de type Newton-Raphson.

D'après les propriétés du maximum de vraisemblance, la matrice de variance de l'estimateur β est donnée par l'inverse de la matrice d'information de Fisher. Ainsi :

$$\widehat{V}(\widehat{\beta}) = \left[\frac{\partial^2 l(\widehat{\beta})}{\partial \beta^2} \right]^{-1} = (\widetilde{X}' \widehat{V} \widetilde{X})^{-1}$$

Où $\widetilde{X}$ est la matrice $n \times (p+1)$ dont les lignes sont composés des $\widetilde{x_t}$ et $\widehat{V}$ est la matrice diagonale $n \times n$ des $\pi(x_i(1 - \pi(x_i))$.

5.6.2 Estimation des odds-ratio

Dans le cas d'une seule variable explicative X, on a

$$\begin{aligned}\ln odds-ratio(x_i,x_j) &= \ln \frac{odds(x_i)}{odds(x_j)} \\ &= logit(\pi(x_i)) - logit(\pi(x_j)) \\ &= \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_j) \\ &= \beta_1 (x_i - x_j)\end{aligned}$$

D'où

$$\widehat{odds-ration} = \exp\left(\widehat{\beta}(x_i - x_j)\right)$$

5.6.3 Redressement dans le cas d'une modalité rare

Nous avons supposé que l'échantillon utilisé pour l'estimation respectait les proportions réelles des deux modalités (échantillonnage simple classique). Or il est très fréquent en pratique, lorsqu'une des deux modalités est rare (présence d'une maladie, client à risque...), d'utiliser un échantillonnage stratifié : on sur-représente artificiellement dans l'échantillon la modalité rare.

Cette modification du schéma d'échantillonnage n'a un impact que sur l'estimation de β_0, qu'il suffit alors de redresser en ajoutant le terme

$$\ln \frac{p_0}{p_1}$$

Où p_0 et p_1 sont les taux de sondage des modalités $Y = 0$ et $Y = 1$(p_0 est donc le rapport de la probabilité d'avoir $Y = 0$ aprés ré-échantillonnage sur cette même probabilité dans la population initiale).

5.7 Prévisions

5.7.1 Classement d'une nouvelle observation

Pour une nouvelle observation x^*, on cherche à prédire y^*. Il existe plusieurs façons d'effectuer la prédiction.

La règle du maximum à postériori (MAP) consiste à affecter l'observation à la classe la plus probable : on prédit donc la valeur de y par la modalité k maximisant la probabilité $P(Y = k|X = x_i;\widehat{\beta})$:

$$\widehat{y}^*_{MAP} = P(Y = k|X = x_i;\widehat{\beta})$$

Puisqu'on est en présence de deux classes, une observation sera classée dans la classe $Y = 1$ si sa probabilité d'être dans cette classe est supérieur à $1/2$. Or, ce choix est totalement arbitraire et peut être remis en cause, notamment lorsque les risques encourus en cas de mauvais classement ne sont pas symétriques (coûte-t-il aussi cher d'accepter un mauvais

client que de ne pas en accepter un bon ?). On définira plus généralement la prédiction, ou règle de classement, au seuil s de la façon suivante :

$$\widehat{y}_s^* = \begin{cases} 1 & si \quad P(Y=1|X=x_i;\widehat{\beta}) \geq s \\ 0 & sinon \end{cases}$$

5.7.2 Tableau de classement ou matrice de confusion

Le résultat d'un procédé de classification est souvent représenté sous la forme d'un tableau de classement (ou matrice de confusion) obtenu en appliquant la méthode de classification sur des observations pour lesquelles la variable Y (i.e. la classe d'appartenance) est connue et en comparant aux classes prédites :

		Prédit		Total
		$Y=0$	$Y=1$	
Réel	$Y=0$	VN	FP	N
	$Y=1$	FN	VP	P
Total		$\widehat{N}$	$\widehat{P}$	n

Matrice de confusion contenant les effectifs de vrais négatifs (VN), vrais positifs (VP), faux négatifs(FN) et faux positifs (FP)

Dans ce tableau figurent les effectifs des observations en fonction de leur classe réelle et de la prédiction de celle-ci. On parle parfois d'observations classées comme positives lorsqu'elles ont la modalité 1 de Y (car bien souvent on associe à la modalité Y = 1 le caractère que l'on cherche à détecter : maladie, achat...), et négatives dans le cas contraire. Avec ces appellations, le contenu des cases du tableau peut être décrit de la faqon suivante :

- vrai négatif (VN) : nombre d'observations pour lesquelles la modalité 0 de Y a correctement été prédite,
- vrai positif (VP) : nombre d'observations pour lesquelles la modalité 1 de Y a correctement été prédite,
- faux négatif (FN) : nombre d'observations détectées à tort comme négatives,
- faux positifs (FP) : nombre d'observations détectées à tort comme positives,
- N,P,$\widehat{N}$ et $\widehat{P}$ respectivement les nombres de négatif et positif réels et prédits. En général, les fréquences sous forme de pourcentage figurent également dans ce type de tableau.

Sensibilité et Spécificité On appelle *sensibilité* du modèle le pourcentage de vrais positifs, et *spécificité* le pourcentage de vrais négatifs.

5.8 Tests, intervalles de confiance

Nous présentons ici les tests permettant d'évaluer l'apport des différentes variables explicatives, ainsi que des intervalles de confiance, notamment sur les odds-ratio, utilisés dans l'interprétation du modèle logistique.

5.8.1 Test sur β_j

On cherche à tester si une composante du paramètre est nulle :

$$H_0 : \beta_j = 0 \qquad contre \qquad H_1 : \beta_j \neq 0$$

Plusieurs tests sont disponibles :

- ***Le test du rapport des vraisemblances maximales***
  Sous H_0 :
  $$-2\ln \frac{\max_\beta L_{H_0}(\beta)}{\max_\beta L_{H_1}(\beta)} \to \chi^2_1$$
  Où L_{H_0} et L_{H_1} sont respectivement les vraisemblances du modèle avec et sans la variable X_j.

- ***Le test de Wald***
  Sous H_0 :
  $$\frac{\widehat{\beta_j}^2}{\widehat{\sigma_j}^2} \to \chi^2_1$$
  Où $\widehat{\sigma_j}^2$ est la variance de l'estimateur de β_j

- ***Le test de score***
  Sous H_0 :
  $$U(\widehat{\beta}_{H_0})'\widehat{V}(\widehat{\beta}_{H_0})U(\widehat{\beta}_{H_0}) \to \chi^2_1$$
  Où $\widehat{V}(\widehat{\beta}_{H_0})$ est l'inverse de la matrice d'information de Fisher, et $U(\widehat{\beta}_{H_0})$ est le vecteur des dérivées partielles de la log-vraisemblance estimée sous H_0.

Pour tous ces tests, on rejettera l'hypothèse de nullité du coefficient β_j si la statistique du test est supérieure au quantile $\chi^2_{1,1-\alpha}$

Remarque 5.2 .

- Si on conclut à la nullité d'un coefficient, tous les autres coefficients doivent être ré-estimés.
- Bien souvent, le test du rapport des vraisemblances est le plus puissant, mais nécessite l'estimation de β sous H_0, ce qui n'est pas le cas pour le test de Wald.

5.8.2 Intervalle de confiance

Sachant que $\widehat{\beta}_j$ est asymptotiquement distribué suivant une loi normale, centrée en $\widehat{\beta}$, et de variance donnée par la matrice d'information de Fisher, il est facile d'en déduire des intervalles de confiance asymptotiques sur les $\widehat{\beta}$.
En pratique, ces intervalles de confiance ne sont que peu souvent utilisés car les $\widehat{\beta}$ ne sont que rarement interprétés, au contraire des odds-ratio. Les intervalles de confiance sur les odds-ratio sont construits à partir de résultats sur la normalité asymptotique du logarithme

d'un odds-ratio.
Un intervalle de confiance sur un odds-ratio qui contient la valeur 1 ne permettra pas de conclure à un effet quelconque de la variable en question.

5.9 Sélection et validation de modèles

1. Sélection : Etant donnée M modèles $M_1, ...M_M$, comment choisir le "meilleur" à partir de l'échantillon dont on dispose.

2. Validation : Est-ce-que le modèle sélectionné M_0 est bon ? En statistique cette question peut être vue de différentes façons :

- Est-ce-que la qualité d'ajustement globale est satisfaisante : le modèle décrit-il bien les valeurs observées ?
  - Ce type de question fait l'objet des tests d'ajustement ou d'équation (goodness of fit).
  - L'ajustement peut être aussi regardé observation par observation (individus aberrants) par des méthodes graphiques (analyse des résidus) ou analytiques.
- Est ce que les hypothèses sont vérifiées ? Les méthodes sont essentiellement graphiques (analyse des résidus).
- L'influence sur l'estimation des points peut être aussi envisagée (distance de Cook, robustesse).

Dans ce chapitre nous allons traiter ces questions à travers l'exemple du modèle logistique. Mais l'ensemble des méthodes que nous présenterons peuvent s'étendre à d'autres problématiques de sélection-validation de modéles.

5.9.1 Sélection ou choix de modéle

Pour la régression logistique, sélectionner un modéle revient à choisir les variables (interactions inclues) qui vont constituer le modèle. On se place dans le cas où on dispose d'un certain nombre de modèles, et on se pose le problème de chercher le meilleur.

5.9.1.1 Un outil spécifique : la déviance

Comme la vraisemblance n'est jamais à la même échelle (cela dépend des données), il n'est pas facile d'avoir une idée de la qualité d'ajustement en regardant la vraisemblance. Pour cela, un outil spécifique est introduit : la déviance. Elle compare la vraisemblance obtenue à celle que l'on obtiendrait dans un modèle parfait : *le modèle saturé*. Elle est définie par :

$$D = 2(L_{sature} - L(\beta)) \geq 0$$

La déviance est égal à 2 fois une différence de vraisemblance. Elle constitue un écart en terme de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré : Dans le modèle saturé, on considère que la prévision est parfaite, c'est à dire que les valeurs prédites sont égales aux valeurs observées. On rappelle que dans le cas où

il n'y a pas de répétitions sur les x_i la log-vraisemblance du modèle logistique est donnée par

$$L(\beta) = \log\left\{\prod_{i=1}^{n} \mathbf{P}(Y = y_i \mid X = x_i)\right\} = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i).$$

Pour le modéle saturé, il n'existe aucune incertitude et la probabilité estimée par le modèle au point $X = x_i$ est donc 1 pour le groupe observé et 0 sinon :

$$\mathbf{P}(Y = j \mid X = x_i) = \begin{cases} 1 & si \quad y_i = j \\ 0 & sinon. \end{cases}$$

Ou encore $\mathbf{P}(Y = y_i \mid X = x_i) = 1$.
Par conséquent $L_{sature} = 0$ et la déviance est égale à deux fois l'opposé de la log-vraisemblance.

Remarque 5.3 Si maintenant plusieurs observations sont effectuées au même point du design, les données étaient alors présentées sous une forme dite binomiale. La log vraisemblance du modèle logistique s'écrit :

$$L(\beta) = \log\left\{\prod_{i=1}^{n} \mathbf{P}(T = t_{ti} \mid X = x_i)\right\} = \sum_{i=1}^{n} \log\begin{pmatrix} n_i \\ t_i \end{pmatrix} + \sum_{i=1}^{n} n_i \{\overline{y}_i \log(p_i) + (1 - \overline{y}_i)\log(1 - p_i)\}$$

où n_i est le nombre d'observations au point x_i et t_i est le nombre de succès associé.
Dans ce contexte le modèle saturé sera tel que :

$$\mathbf{P}(Y = y_i \mid X = x_i) = \overline{y}_i$$

On aura donc :

$$L_{sature} = \sum_{i=1}^{n} \log\begin{pmatrix} n_i \\ t_i \end{pmatrix} + \sum_{i=1}^{n} n_i \{\overline{y}_i \log \overline{y}_i + (1 - \overline{y}_i)\log(1 - \overline{y}_i)\}$$

La déviance sera alors égale à

$$D = 2\sum_{i=1}^{n} n_i \left(\overline{y}_i \log\frac{\overline{y}_i}{p_i} + (1 - \overline{y}_i)\log\frac{1 - \overline{y}_i}{1 - p_i}\right).$$

5.9.1.2 Test de déviance entre 2 modèles emboîtés

Par définition un modèle est dit emboîté dans un autre plus général (ou plus grand) lorsqu'il est un cas particulier de ce modèle plus général.

Exemple 5.2 Dans le cas de la régression simple, le modèle

$$y = \beta_0 + \beta_1 x_1 + \varepsilon,$$

est un cas particulier du modèle

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

En effet il suffit de poser que $\beta_2 = 0$ dans le second modèle pour retrouver le premier. Notons les estimations dans le modèle 1 $(\widehat{\beta}_0^{(1)}, \widehat{\beta}_1^{(1)})$ et dans le modèle $2(\widehat{\beta}_0^{(2)}, \widehat{\beta}_1^{(2)}, \widehat{\beta}_2^{(2)})$. En général nous avons $\widehat{\beta}_0^{(1)} \neq \widehat{\beta}_0^{(2)}$ et $\widehat{\beta}_1^{(1)} \neq \widehat{\beta}_1^{(2)}$.

Dans le cas d'un modèle logistique binaire, cela est identique

$$(\mathbf{P}(Y = 1 \mid X = x) = beta_0 + \beta_1 x_1 + \beta_2 x_2$$

et

$$(\mathbf{P}(Y = 1 \mid X = x) = beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

sont emboîtés l'un dans l'autre. Pour comparer deux modèles emboîtés $M_1 \subset M_2$ nous allons comparer leur déviance D_1 et D_2. On a alors deux cas :

- La différence est grande $\rightarrow$ le fait de passer d'un modèle simple (petit) à un modèle plus complexe (plus génvral ou plus grand) a donc apporté un écart de déviance significatif $\rightarrow$ le modèle plus général est acceptable.
- La différence est faible $\rightarrow$ le modèle simple et celui plus complexe sont voisins et par souci de parcimonie le modèle simple est conservé.

Il nous faut bien entendu déterminer un seuil à partir duquel on pourra dire que la différence de déviance est petite ou grande. Pour se faire, on construit un test dans lequel nous allons chercher la loi de la différence de déviance sous H_0 (l'hypothèse selon laquelle le modèle simple est vrai).

Sous des hypothèses techniques $\Delta D = D_1 - D_2 = D_{petit} - D_{grand}$ suit une loi du χ^2 à $p_2 - p_1$ degrés de liberté où p_1 est le nombre de paramètres du modèle simple et p_2 celui du modèle complexe. Le test se déroule alors de la manière classique

1. Les hypothèses sont fixées
   - H_0 le modèle simple à p_1 paramètres est adéquat ;
   - H_1 le modèle complexe à p_2 paramètres est adéquat.
2. α est choisi (en général 5% ou 10%).
3. L'observation de ΔD est calculée, notons la ΔD_{obs}
4. Calcul du quantile de niveau $(1-\alpha)$ de la loi du $\chi^2(p_2 - p_1)$, noté $q_{1-\alpha}(p_2 - p_1)$.
   - Si $\Delta D_{obs} > q_{1-\alpha}(p_2 - p_1)$ alors H_0 est repoussé au profit de H_1, le modèle éré n'est pas adéquat.
   - Si $\Delta D_{obs} \leq q_{1-\alpha}(p_2 - p_1)$ alors H_0 est conservé, le modèle considéré est adéquat.

5.9.2 Critère de choix de modèles

Le test que nous venons d'étudier permet de sélectionner un modèle parmi deux modèles emboîtés. Or en régression logistique, nous avons vu qu'à partir de p variables explicatives, nous pouvions construire une grand nombre de modèles logistiques, qui ne sont pas forcément emboîtés. L'utilisation d'un simple test de déviance se révèle alors insuffisante. On a alors recours à des critères de choix de modèles qui permettent de comparer des modèles qui ne sont pas forcément emboîtés les uns dans les autres.
Les critères **AIC** et **BIC** sont les plus utilisés. Ces critères sont basés sur la philosophie suivante : plus la vraisemblance est grande, plus grande est donc la log-vraisemblance L et meilleur est le modèle. Cependant la vraisemblance augmente avec la complexité du modèle, et choisir le modèle qui maximise la vraisemblance revient à choisir le modèle saturé. Ce modèle est clairement surparamétré, on dit qu'il "sur-ajuste" les données (overfitting). Sur l'exemple de la Figure 5.2, nous avons simulé un échantillon de taille 100

suivant :

$$X_i \sim \mathcal{N}(0,1), \qquad U_i \sim \mathcal{U}[0,1], \qquad Y_i = \begin{cases} \mathbf{1}_{U_i \leq 0.25} & si \quad X_i \leq 0 \\ \mathbf{1}_{U_i \geq 0.25} & si \quad X_i \geq 0 \end{cases}$$

Dit autrement, environ $3/4$ des labels valent 0 pour les valeurs de X_i négatives et 1 pour les valeurs positives. De manière évidente, le modéle saturé ajuste parfaitement les observations. Nous voyons cependant qu'il est difficile, pour ne pas dire impossible à utiliser dans un contexte de prévision. De plus le modèle saturé possède ici $n = 100$ paramètres tandis que le modèle logistique n'en possède que 2. Ce qui est nettement plus avantageux pour expliquer Y.
Pour choisir des modèles plus parcimonieux, une stratégie consiste à pénaliser la vraisemblance par une fonction du nombre de paramètres.

- Par définition l'AIC (*Akaike Informative Criterion*) pour un modèle à p paramètres est
$$AIC = -2L + 2p.$$
- Le critère de choix de modèle le BIC (*Bayesian Informative Criterion*) pour un modèle à p paramètres estimé sur n observations est défini par
$$BIC = -2L + p\log(n).$$

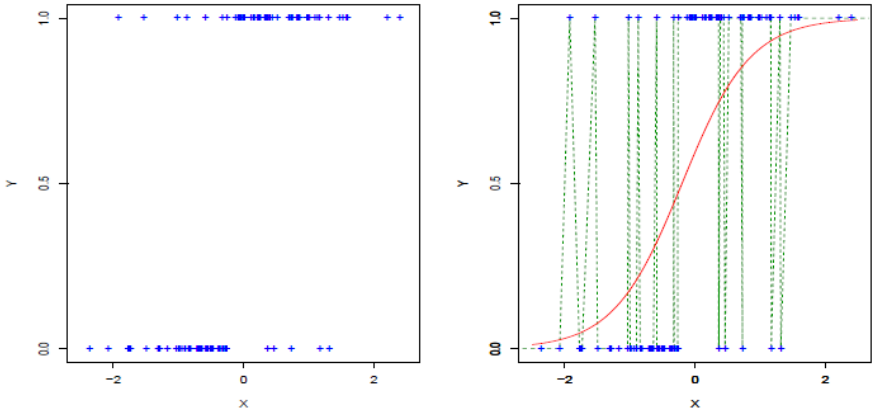

FIGURE 5.2 – Gauche : Représentation des observations . Droite : Tracé des modèles saturés (pointillés) et logistique (trait plein).

On choisira ainsi le modèle qui possède le plus petit AIC ou BIC. L'utilisation de ces critères est simple. Pour chaque modèle concurrent le critère de choix de modèle est calculé et le modèle qui présente le plus faible est sélectionné.

5.9.2.1 Algorithme de sélection de variables :

Comme en régression multiple, il existe des algorithmes de sélection (forward, backward, stepwise...) dont le principe est à chaque étape de comparer un modèle avec un

sous-modèle et d'évaluer l'apport de termes supplémentaires.
Le critère utilisé est généralement la statistique issue des tests de Wald ou du rapport des vraisemblances maximales.

5.9.3 Validation du modèle

5.9.3.1 Test d'adéquation par la déviance

Ce test permet de valider un modèle à p paramètres. Les hypothèses nulle et alternatives sont :

- H_0 le modèle considéré à p paramètres est adéquat ;
- H_1 le modèle considéré à p paramètres n'est pas adéquat.

Ici, nous allons comparer le modèle saturé au modèle considéré au moyen de la déviance. Nous savons que

- si la déviance est grande, alors le modèle considéré est loin du modèle saturé et que par conséquent il n'ajuste pas bien les données ;
- Par contre si la déviance est proche de 0, le modèle considéré sera adéquat.

La déviance est en fait le test de rapport de vraisemblance et sous des hypothèses techniques, D suit donc une loi du $\chi^2(n-p)$ degrés de liberté, où p est le nombre de paramètres du modèle et n le nombre de point du design, ce qui est, sauf répétition, le nombre d'observations. Le test se déroule alors de la manière classique :

1. Les hypothèses sont fixées
   - H_0 le modèle considéré à p paramètres est adéquat
   - H_1 le modèle considéré à p paramètres n'est pas adéquat
2. α est choisi (en général 5% ou 10%)
3. L'observation de D est calculée, notons la D_{obs}
4. Calcul du quantile de niveau $(1-\alpha)$ de la loi du $\chi^2(n-p)$, noté $q_{1-\alpha}(n-p)$.
   - Si $D_{obs} > q_{1-\alpha}(n-p)$ alors H_0 est repoussé au profit de H_1, le modèle considéré n'est pas adéquat.
   - Si $D_{obs} \leq q_{1-\alpha}(n-p)$ alors H_0 est conservé ,le modèle considéré est adéquat.

5.9.3.2 Test d'Hosmer Lemershow

Ce test [15] permet de vérifier l'adéquation d'un modèle en présence de données individuelles, il relève à peu près de la même logique que le diagramme de fiabilité. A la différence qu'au lieu de se baser simplement sur une impression visuelle, on extrait du tableau de calcul un indicateur statistique qui permet de quantifier la qualité des estimations $\widehat{\pi}(x)$.
Concrètement, nous procédons de la manière suivante :

1. Les probabilités $\widehat{\pi}_i$ sont ordonnées par ordre croissant ($\widehat{\pi}_i$ est la probabilité $P(Y = 1 \mid X = x_i)$ estimée par le modèle) ;
2. Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand).

On note

- m_k les effectifs du groupe k ;
- g_k le nombre de succès $(Y = 1)$ observé dans le groupe k ;

- $\mu_k$la moyenne des $\widehat{\pi}_i$ dans le groupe k.

La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(g_k - m_k\mu_k)^2}{m_k\mu_k(1-\mu_k)}$$

Le test se conduit de manière identique au test de déviance, la statistique C^2 suit approximativement un $\chi^2(K-1)$ degrés de liberté.Cette approximation ayant été validée uniquement par simulation, il semble donc important de ne pas appliquer trop strictement la procédure de test, mais plutôt de la considérer comme une indication.

5.9.4 Analyse des résidus

5.9.4.1 Les différents types de résidus

A l'image de la régression plusieurs types de résidus sont proposés par les logiciels. Le premier, le plus simple à calculer est tout simplement $Y_i - \widehat{\pi}_i$. Ces résidus sont appelés résidus bruts. Ils permettent de mesurer l'ajustement du modèle sur chaque observations. Ces résidus n'ayant pas ma même variance, ils sont difficiles à comparer. En effet, on rappelle que $\mathbf{V}(Y|X = x_i) = \pi_i(1-\pi_i)$, et par conséquent, de tels résidus risquent d'être pour des valeurs de pi proches de 1/2. Un moyen de pallier à cette difficulté est de considérer **les résidus de Pearson**

$$r_i = \varepsilon_i = \frac{Y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}} \tag{5.7}$$

Par définition on standardise les résidus par la variance théorique de Y_i qui prend comme valeur 0 ou 1. La variance théorique est donc celle d'une loi de Bernouilli $\pi_i(1-\pi_i)$. Ce n'est pas la variance de l'estimation $\widehat{\pi}$ qui est un estimateur donc aléatoire. On note

$$\begin{cases} \varepsilon_i = Y_i - \pi_i \\ \widehat{\varepsilon}_i = Y_i - \widehat{\pi}_i \end{cases}$$

Pour essayer d'obtenir des résidus de même variance approximative (**standardisés**)

$$\widehat{\varepsilon}_i = \frac{Y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}(1-\widehat{\pi})(1-h_{ii})}},$$

Les résidus de déviance sont définis par

$$d_i = \widehat{\varepsilon}_i = (Y_i - \widehat{\pi}_i)\sqrt{2(l_{sature}(Y_i) - l(Y_i,\beta))},$$

où $l(Y_i,\beta)$ est la log-vraisemblance associée à l'observation Y_i (et non pas toutes les observations) et $l_{sature}(Y_i)$ son homologue pour le modèle saturé. Cette définition est moins naturelle. Là encore pour tenir compte de la variabilité ces résidus sont standardisés :

$$\widehat{\varepsilon}_i = (Y_i - \widehat{\pi}_i)\sqrt{\frac{2(l_{sature}(Y_i) - l(Y_i,\beta,\phi))}{1-h_{ii}}}$$

Ces deux types de résidus de déviance sont ceux qui sont en général conseillés.

Les résidus partiels Les résidus partiels sont définis par

$$\widehat{\varepsilon}^{P}_{.j} = \frac{Y_i - \widehat{\pi}_i}{\widehat{\pi}_i(1-\widehat{\pi}_i)} + \widehat{\beta}_j X_{.j}$$

5.10 Un outil d'interprétation : la courbe ROC

Nous avons défini précédemment les notions de sensibilité (pourcentage de vrais positifs) et spécificité (pourcentage de vrai négatif). La courbe ROC *Receiver Operaor Characteristic curve* donne l'évolution du taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1-spécificité) lorsqu'on fait bouger le seuil s utilisé pour la prédiction.

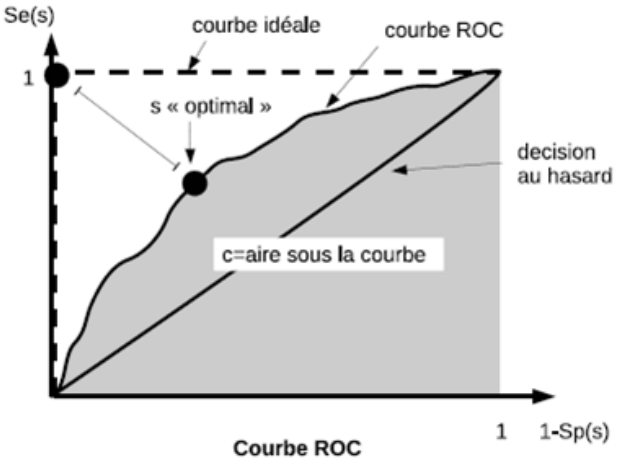

FIGURE 5.3 – courbe ROC

Cette courbe permet de voir l'évolution des sensibilité et spécificité en fonction du seuil s choisi. Le praticien pourra alors choisir le seuil :

- A la main en fonction d'une sensibilité ou spécificité souhaitée,
- De façon à minimiser l'erreur totale de classement (sans différencier les FP et FN), c'est − à − dire le seuil s minimisant :

  $$p_0 1 - Se(s)\,) + p_1(\,1 - Sp(s)\,)$$

  Où $Se(s)$ $Sp(s)$ sont les sensibilité et spécificité (en fonction du seuil s), et p_0 et p_1 sont les proportions de négatifs et de positifs dans la population totale,
- En cherchant à être le plus près possible du point idéal de coordonnées $(0,1)$,$(Se = Sp = 1)$ c'est-à-dire en minimisant :

  $$(1 - Se(s)\,)^2 + \big(\,1 - Sp(s)\,\big)^2$$

La courbe ROC permet également d'évaluer la qualité du modèle. Pour cela, on calcule l'aire sous cette courbe, notée AUC (Area Under Curve) :

$$AUC = \int_0^1 Se(s)d(1 - Sp(s)\,)$$

Le meilleur modèle sera celui qui se rapprochera le plus de l'AUC maximale égale à 1. Cette aire correspond à la probabilité de détecter un positif d'un négatif.

5.11 Le modèle logistique polytomique et ordinal

Le modèle logistique présenté précédemment se généralise au cas d'une variable Y à K modalités ($K > 2$).
Lorsque ces dernières sont ordonnées on parle de régression logistique ordinale.
Notons $\pi_k(x) = P(Y = k \mid X = x)$. Dans cette situation, on se fixe une modalité de référence($Y = K$ par exemple) et on réalise $K-1$ régressions logistiques de $\pi_k(x)$ versus $\pi_K(x)$:

$$\ln \frac{\pi_k(x)}{pi_K(x)} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j \qquad \forall 1 \leq k \leq K-1.$$

Cette procédure ne dépend pas du choix du groupe de référence (dans les logiciels le groupe de référence est généralement soit le premier soit le K^{ime}).
Lorsque la variable est ordinale, on modélise généralement des logits cumulatifs :

$$\ln \frac{\pi_{k+1}(x) + ... + \pi_K(x)}{\pi_1(x) + ... + \pi_k(x)} = \sum_{j=1}^{p} \beta_{jk} x_j \qquad \forall 1 \leq k \leq K-1.$$

Ce dernier modèle comportant un grand nombre de paramètres, les β_{jk} sont souvent supposés constants par classe β_{jk} $\forall 1 \leq k \leq K-1$.

Chapitre 6

Application

Sommaire

6.1 Introduction

Les aspects théoriques ayant été abordés, nous allons à présent mettre en avant un exemple de modélisation. Dans un premier temps, nous présenterons brièvement le jeu de données, ainsi que la méthodologie de construction de la modélisation qui sera retenue et les objectifs de cette dernière. Puis nous nous intéresserons plus en détails à la modélisation de la durée de la première survenue d'un sinistre. Afin de modéliser le processus de sinistralité, nous désirons savoir quel est le déroulement "type" de la durée de vie d'un sinistre. Pour cela nous cherchons à décrire sa manière d'évoluer, et quels pourraient être les facteurs explicatifs de cette évolution. Nous souhaitons ici traiter la durée de vie des sinistres de deux manières, à traves les modèles linéaires généralisés et les modèles de survie.

Nous faisons remarquer que l'ensemble des résultats mis en avant ici ont été obtenus à partir du logiciel R.

6.2 Description des données

Ce chapitre porte sur des données issues d'un portefeuille d'assurance automobile en Algérie, nous disposons d'une table prenant l'ensemble des contrats couvrant d'au moins un jour de garantie durant l'année 2011, l'ensemble des sinistres survenus, et des informations sur les conducteurs assurés relatives à n=6412 police observées . Les variables reprises dans le fichier son données au Tableau 6.1. Compte tenu de l'objectif de notre sujet, il convient de mentionner que nous nous intéresserons particulièrement à la durée de survenue d'un sinistre pour les différents contrats souscrits. Ainsi pour un véhicule voyant plus d'une fois la survenue d'un sinistre pendant la même période de garantie, on s'intéressera uniquement au premier accident de leurs différents sinistres. En premier lieu, précisons quelques points importants. A l'exception des variables DSURV et CTOT décrivant la sinistralité, il s'agit toutes de variables connues à priori de l'assureur (c'est-à-dire qu'il peut se servir de ces variables pour personnaliser le montant de la prime réclamé à l'assuré pour la couverture du risque). Parmi les variables explicatives disponibles au tableau nous distinguons différent types :

i Celles relatives au preneur d'assurance : AGEC, SEXE
ii Celles relatives au véhicule assuré : AGEV, PUISVEHI
iii Celle relative à la couverture pour la quelle le preneur a opté : PRIME

On s'intéresse à la modélisation de la variable durée de survenue d'un sinistre de manière plus précise le premier sinistre qui a eu lieu durant l'année 2011, ainsi on se limite à un sous échantillon de nos données, après avoir éliminer les sinistres répétés d'un client souscrit et vérifier la cohérence des chiffres données , notre échantillon d'étude est de taille 3394 sinistres observés.

Avant d'entamer l'étude de la durée DSURV de l'échantillon réduit (c'est-à-dire celui qui porte sur l'observation du premier sinistre), il est nécessaire de bien connaitre les données sur les quels on travaille. Il faut dés lors décrire en détail les variables disponibles qui peuvent nous servir à la modélisation en les considérant comme variables explicatives

Variable	Description
Numpolice	Numéro de la police d'assurance
DCOUV	Durée de couverture (en jours)Date de résiliation - Date de souscription
DSURV	Durée de survenue d'un sinistre (en jours) =Date de survenue - Date de souscription
AGEV	Age du véhicule assuré
AGEC	Age du souscripteur
SEXE	Sexe du souscripteur
PRIME	Prime versée
CTOT	Cout total du sinistre
PUISVEH	Puissance du véhicule assuré

TABLE 6.1 – Variables comprises dans la base des données

de la variable d'intérêt DSURV .

Nous travaillions ici sur un fichier polices. Un tel fichier compte autant de lignes que de polices sélectionnées pour notre échantillon du portefeuille global durant l'année 2011. Il résume les informations disponibles en début de période et décrit la sinistralité relative à ceux-ci.

6.2.1 Variables décrivant la sinistralité :Le nombre des sinistres NSIN

Il s'agit du nombre de sinistres déclarés par l'assuré à la compagnie, et donc pas du nombre de sinistres causés par l'assuré sur l'année. L'assuré peut en effet estimer avoir l'intérêt à dédommager lui-même le tiers lésé en cas de préjudice mineur.
En RC automobile, NSIN méritent une intention particulière car les coûts des sinistres ne se prêtent souvent pas à une segmentation poussée. De plus, NSIN jouera un rôle central dans la personnalisation à posteriori des montants des primes. Le gros avantage de cette variable est d'être généralement connu avec précision par la compagnie.

Nombre de sinistre(k)	Nombre de polices obs.
1	3394
2	1815
3	669
4	340
5	70
6	96
7	28

TABLE 6.2 – Sinistralité observée dans le portefeuille

Le tableau nous apprend que le nombre maximum de sinistres déclarés par assuré

vaut 7. Plus précisément, 3394 (soit 52.95%) ont déclaré un seul sinistre, 1815 (28.3%) et ont déclaré 2, et 28 (0.43%) et ont déclaré 7, au cours de l'année 2011.

6.2.2 La mesure de l'exposition au risque : La variable DCOUV

Il s'agit du nombre de jours où la police a été en vigueur durant l'année 2011. On peut s'en servir pour mesurer l'exposition au risque. Elle permettra de tenir compte qu'un sinistre déclaré par une police en vigueur durant un mois est un plus mauvais signe pour l'assureur qu'un sinistre relatif à une police en vigueur toute l'année.

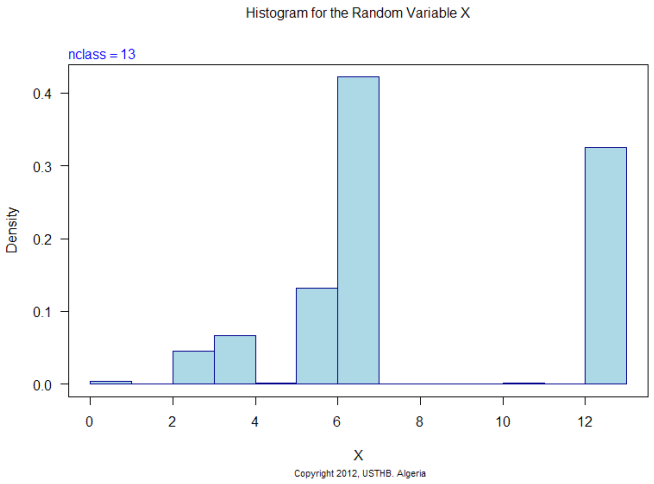

FIGURE 6.1 – Durée de couverture

Dans notre échantillon, la durée moyenne de couverture est de 230 jours, quelques polices n'ont pas été en vigueur que pendant une seule journée. La Figure 6.1 donne une idée des périodes de couverture des différentes polices sinistrées pour la première fois en 2011. On constate une majorité de polices couvertes durant une période de 6 mois et une grande partie de police couverte toute l'année. Les polices dont l'exposition au risque est inférieure à 6 ou 12 mois sont les nouvelles affaires et les résiliations.

6.2.3 Caractéristique du preneur d'assurance : Variable AGEC

Il s'agit d'une variable quantitative à valeurs entières donnant l'âge du preneur d'assurance (en années révolues) au premier janvier de l'année 2011.Examinons à présent la structure dans notre échantillon. Celle-ci est décrite à la Figure 6.2. On constate clairement une sursinitralité des jeunes et moyens conducteurs. Avec l'âge la sinistralité a tendance à diminuer chez les plus âgés.

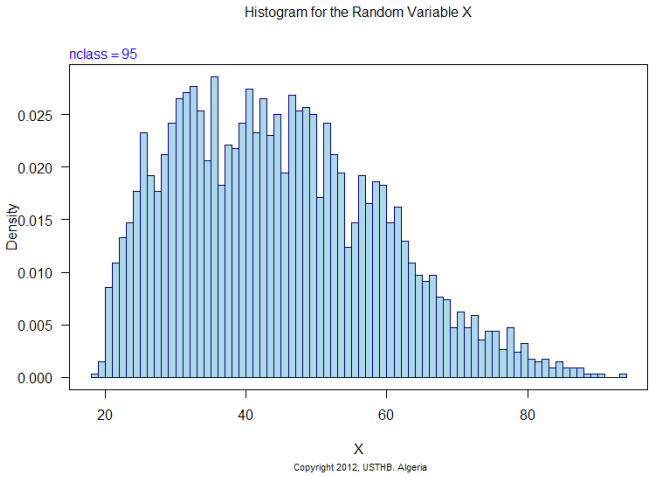

FIGURE 6.2 – Age du souscripteur

6.3 Modélisation de la durée DSURV

Variable à expliquer Nous avons ici la variable quantitative durée de survenue du premier sinistre calculée en jours qui représente la différence entre la date de survenue et la date de souscription pendant l'année 2011.

Variables explicatives Ce sont pour la plupart les variables décrivant le risque automobile, dans notre cas nous avons utilisé quatre variables explicatives : AGEC, SEXE, AGEV et PUISVEH.

6.3.1 Analyse graphique de variable DSURV

Rappel (Estimateur à noyau pour la densité)
appelé aussi estimateur de densité de *Parzen-Rosenblatt* qui a été introduit par *Rosenblatt*(1956) et développé par *Parzen*(1962).

$$\widehat{f_n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{6.1}$$

où K est un noyau (Kernel en anglais), et h_n fenêtre ou paramètre de lissage.

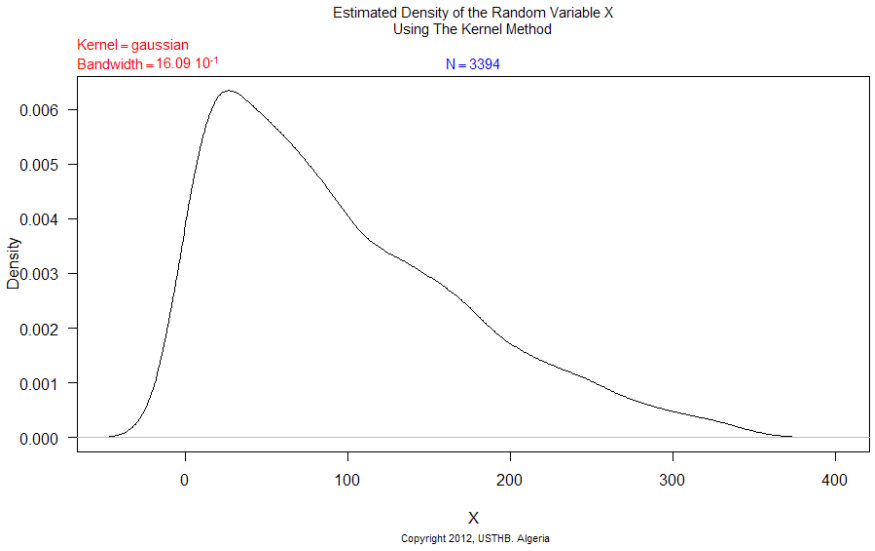

FIGURE 6.3 – Estimation graphique (DSURV)

La visualisation graphique des observations de l'échantillon nous permettra de voir la forme de la distribution de la durée (en jours). La Figure 6.3 montre une estimation graphique de la variable DSURV par la méthode Kernel 6.1. Un examen rapide de ce graphique permet de constater que la durée est caractérisée par une queue des valeurs peuvent être évalués comme valeurs extrêmes par rapport aux pics enregistrés au début de période.

Dans le but de décider s'il est vraisemblable que la modélisation de la variable durée de première survenue soit adéquate aux données réelles un traitement particulier aux durées extrêmes peut être de grande utilité.

Ces constations montrent l'intérêt de la prise en compte des valeurs extrêmes dans la modélisation des durées des sinistres et de la détermination des paramètres à fixer par l'assureur. En effet ; par exemple l'omission ou l'élimination des observations extrêmes pour une raison ou une autre tend à sous estimer la prime pure qui aura pour effet un manque à gagner et même des pertes pour l'assureur. De plus la prise en compte des coûts extrêmes montre la nécessité pour l'assureur de réviser sa logique de segmentation et de classification des risques.

6.3.2 Détermination du seuil et estimation de l'indice de queue

Pour l'estimation d'un seuil au-delà duquel la durée sera jugée extrême, nous avons procédé par la fonction moyenne des excès (MEF), nous avons appliqué la méthode sur notre échantillon.

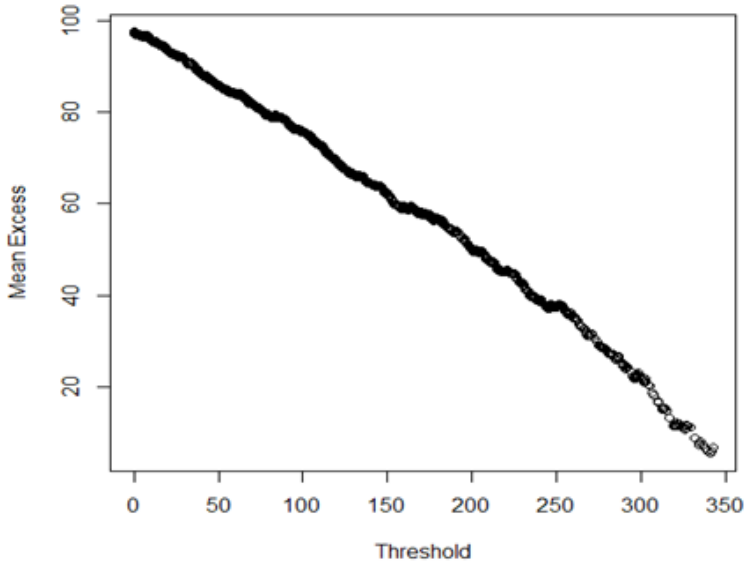

FIGURE 6.4 – Fonction moyenne des excès

La Figure 6.4 présente la FME (fonction moyenne des excès) empirique d'une distribution GPD (Generalized Pareto Distribution) avec un paramètre ξ négatif. On constate que la FME devient stable pour un seuil de l'ordre de 250 jours. Ces valeurs extrêmes nous emmènent à examiner la variable DSURV en tenant compte de ce phénomène.

Rappelons que dans ce chapitre notre but est de proposer un modèle qui s'ajustera le mieux aux données observées ; ceci dans l'intention de mettre en exergue les facteurs décrivant de façon significative la sinistralité. Le but final étant de dire au moyen de ses caractéristiques si le nouveau client est à risque ou non. Un client avec une durée de survie supérieure au seuil τ peut juger comme un bon client, on s'intéresse sur l'autre catégorie de l'échantillon plus précisément sur les clients à un niveau de risque plus élevé qui ont causés un sinistre durant la période allant de 0 à 50 jours, on a choisit de fixé un seuil $\tau = 50$ jours pour segmenté l'échantillon en deux catégories, tel que les clients avec DSURV inférieur à τ nécessitent un traitement particulier. Cette procédure de segmentation nous conduit à un résultat binaire (ou dichotomique) que l'on souhaite mettre en relation avec des variables explicatives.

En général, le résultat d'une observation binaire est appelé "succès" ou "échec". Il est représenté mathématiquement par une variable aléatoire Y telle que $Y = 1$ s'il y a succès et $Y = 0$ s'il y a échec.
Le résultat Y peut dépendre des valeurs assumées par p variables explicatives $X_1, , X_p$ au moment de l'observation et nous souhaitons étudier cette relation.

Application 1 :
Le Modèle de Régression Logistique

Les techniques de régression ordinaire ne sont pas adaptées à ce type d'analyse, on cherche à applique la méthode de régression logistique qui permet de traiter le cas où la variable réponse est de type binaire, et non pas continu comme dans le modèle de régression linéaire. Tout en relaxant certaines hypothèses du modèle de régression multiple, on maintient quand même l'idée d'une relation linéaire entre la réponse et les prédicteurs.

Pour ce but on procède à définir une nouvelle variable Y à deux modalités :

$$Y = \begin{cases} 1, & \text{si la durée DSURV est inférieure à } \tau \\ 0, & \text{sinon} \end{cases}$$

Y est une variable dichotomique (ou binaire) à expliquer, dite aussi dépendante, dont on veut avoir un modèle significatif représentatif de la durée DSURV.

6.4 Modèle de régression logistique pour la durée DSURV

Soit y_i la valeur déduite (calculée) de Y qui correspond à $dsurv_i$ la i^{me} observation de la variable DSURV, $i = 1,2,..,n$ et $n = 3394$. Le modèle de régression logistique repose sur l'hypothèse d'une relation entre la probabilité p de maladie et les variables explicatives $X_1, X_2, , X_p$ de la forme :

$$P(Y = 1 \mid X_1, X_2, ..., X_p) = \pi(x) = \frac{e^{\beta_0 + \Sigma_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \Sigma_{j=1}^{p} \beta_j X_j}}$$

ou, ce qui est équivalent,

$$logit(\pi(x)) = \ln\left(\frac{\pi(x)}{1+\pi(x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Où :

β est un vecteur de coefficients de régression inconnus.
Y est la variable dichotomique $Y \in \{0,1\}$.
$X = \{AGEC, SEXE, AGEV, PUISVEH\}$ vecteur des régresseurs.

Dans le paragraphe précédent nous avons déterminé les facteurs décrivant de façon significative la sinistralité. Cette fois nous allons considérer un modèle multivariables afin d'étudier l'effet conjoint de plusieurs covariables sur la probabilité d'avoir une hausse durée de sécurité ; en essayant d'estimer la contribution pour chaque facteur à l'explication de celle ci. Nous utiliserons pour cela la procédure de sélection pas à pas (implémentée dans le logiciel R) du modéle de régression logistique.

6.4.1 Estimation du modèle

6.4.1.1 Modèle obetenu pas à pas

On commence par le modèle qui contient toutes les covariables ; On èlimine à chaque étape la covariable qui a la plus grande p-value (probabilité sous l'hypothèse nulle de rejeter l'hypothèse nulle), jusqu'à ce que les covariables restantes aient une p-value inférieure à une limite donnée. Ici on la prend égale à 0.1 (test au seuil 10%) On obtient le premier résultat suivant :

```
glm(formula=Y ∼ AGEV + AGEC + PUISVEH + SEXE, family = binomial(link =
"logit"), data = data)
```

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.036951	0.277708	1.933	0.0894 *
AGEC	-0.005039	0.002544	-1.981	0.0476 *
AGEV	0.009138	0.004502	2.030	0.0424 *
SEXE	-0.220614	0.183342	-1.203	0.2289
PUISVEH	-0.054223	0.025353	-2.139	0.0325 *

TABLE 6.3 – Estimations des coefficients du modèle de régression correspondant aux (04) variables exogènes considérées.

```
 Null deviance: 4356.7 on 3393 degrees of freedom
Residual deviance: 4344.2 on 3389 degrees of freedom
AIC: 4354.2
Number of Fisher Scoring iterations: 4
Signification des codes :
NS : non significatif; . significatif à 10% * : significatif à 5%;
*** : très significatif
```

On supprime dans un premier temps la variable SEXE qui a la plus grande p-value et on reprend le procédé.

```
glm(formula = Y ~ AGEC + AGEV + PUISVEH, family = binomial(link =
"logit"), data = data)
```

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.212513	0.185177	-1.963	0.0334 *
AGEC	-0.003748	0.002125	-1.916	0.0610 .
AGEV	0.008849	0.004257	2.176	0.0343 *
PUISVEH	-0.032567	0.02869	-2.023	0.0397 *

TABLE 6.4 – Estimations des coefficients du modèle de régression correspondant au **Modèle 1**.

```
Null deviance: 4356.7 on 3393 degrees of freedom
Residual deviance: 4345.7 on 3390 degrees of freedom
AIC: 4353.7
Number of Fisher Scoring iterations: 4
```

Les résultats de la deuxième étape ont montré que les variables restantes avaient une p-value inférieure au seuil fixé 0.1. On a obtenu alors le modèle comprenant les covariables AGEV, AGEC et PUISVEH. On retient donc au final le **Modèle 1**.

Remarque 6.1 Le **Modèle 1** ci-dessus est admissible car les variables sont toutes significatives au seuil indiqué (10%). Cependant il est important pour nous d'inspecter son graphique de diagnostic.
Nous remarquons ensuite que l'ajustement du modéle logistique nous a fournit le critère AIC , ainsi que les coefficients et écarts types estimés des paramètres associés à chaque occurrence des prédicteurs .Les deux colonnes suivantes sont issues d'un test de student visant à tester l'hypothèse de nullité d'un coefficient. Lorsque la p-value est faible, le coefficient associé est significativement non nul .

6.4.2 Graphique de diagnostic

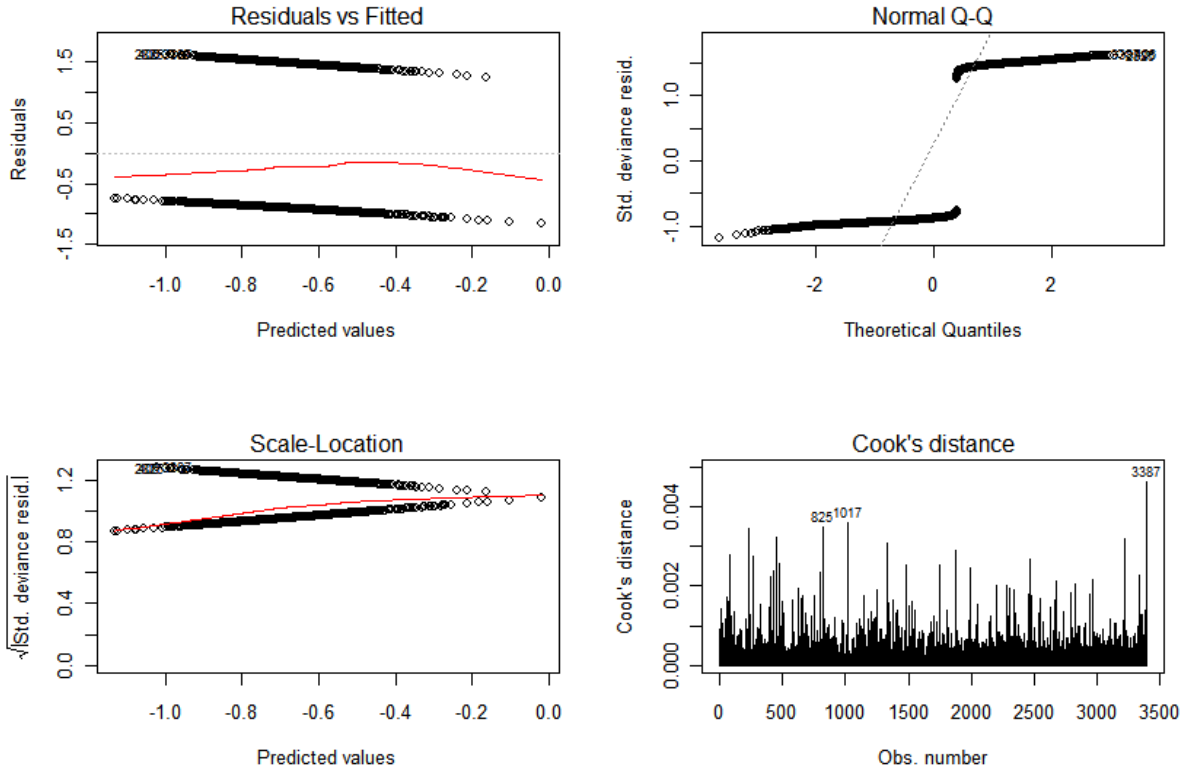

FIGURE 6.5 – Quelques graphes de diagnostic du **Modèle 1**

Le premier graphique (en haut à gauche) est une représentation des résidus en fonction des valeurs prédites, l'absence de tendance significative et la dispersion des points autour de l'ordonnée 0 (comme c'est le cas ici) indique une bonne adéquation du modèle au problème.
Le second graphique permet de contrôler l'adéquation des résidus à une loi normale. Le troisième est une représentation de la racine des résidus (en valeurs absolues) en fonction des valeurs prédites. Comme pour le premier graphique, l'absence de tendance est la preuve d'une bonne adéquation.
Enfin le dernier graphique est celui des distances de Cook. Une distance supérieure à 1 sera considérée comme anormale, nous remarquons que toutes les observations de l'étude sont convenables ici.

Remarque 6.2 Le graphe des coefficients de Cook [16] montre que certaines observations ont des coefficients de Cook importants, notamment les observations 825, 1017 et 3387 qui pourraient être des outliers (valeurs aberrantes). Il convient d'ailleurs de rappeler que le graphe des distances de Cook (Cook's distance plot) mesure pour un individu l'écart entre la valeur observée et celle prédite par le modèle. Une distance trop grande signifie donc que l'ajustement n'est pas correct en ce point.

Essayons maintenant de supprimer les enregistrements 825, 1017 et 3387 de notre base de données.
Après cette suppression, le graphique des distances de Cook montre à nouveau trois outliers : Les enregistrements 233, 3217, et 3384. Ainsi nous supprimons au total six enregistrements de la base de données initiale. Notre base de données comporte désormais 3388 enregistrements. Le modèle ajusté à cette nouvelle base est donné ci-dessous ; appelons le **"Modèle 2"**.

```
glm(formula=Y ∼ AGEV + AGEC + PUISVEH , family = binomial(link =
"logit"), data = data2)
```

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.214216	0.183951	-1.965	0.0244*
AGEC	-0.004897	0.002542	-1.927	0.0540 .
AGEV	0.009594	0.004486	2.138	0.0325 *
PUISVEH	-0.052367	0.025297	-2.070	0.0384 *

TABLE 6.5 – Estimations des coefficients du modèle de régression correspondant au **Modèle 2**.

```
Null deviance: 4352.8 on 3390 degrees of freedom
Residual deviance: 4341.7 on 3387 degrees of freedom
AIC: 4342.9
Number of Fisher Scoring iterations: 4
```

Remarque 6.3 On constate que le modèle obtenu après suppression de quelques valeurs aberrantes (**Modèle 2**) s'ajuste nettement mieux aux données, comparé au **Modèle 1**. On peut le voir en comparant les critères d'information d'Akaike AIC [4] associés aux deux modèles ; l'ajustement étant d'autant plus bon que son AIC est faible. Cet écart considérable entre les deux critères d'information (4353.7 et 4342.9) nous permet de voir l'effet néfaste que pourraient apporter ces valeurs aberrantes dans notre modélisation. On s'intéressera donc pour la suite au **Modèle 2**.

6.4.3 Interprétation des résultats

6.4.3.1 Interprétation des coefficients

Les coefficients du modèle s'interprètent bien :

- Les véhicules d'une forte puissance fiscale sont corrélés négativement avec la sinistralité. Ceci signifie que les véhicules de pareilles puissances fiscales sont les moins à risque, car le signe'-' signifie que cette caractéristique diminue la probabilité d'avoir l'événement "Y=1", par conséquent diminue la probabilité d'être "haut sinistré". Le cofacteur AGEC s'interprète de la même façon.
- On remarque aussi que la variable AGEV est corrélée positivement avec la variable d'intérêt Y ; c'est-à-dire que les anciens véhicules sont à risque. De plus le symbole '$*$' nous montre que le degré de significativité du test de nullité du coefficient associé à cette variable est élevé ; ce qui signifierait qu'une grande majorité des sinistres dont la durée est inférieure à τ sont issus de cette classe de véhicule.

6.4.3.2 Interprétation du terme constant

Quelle signification donner un terme constant β_0 de la régression ? Si toutes les variables explicatives sont nulles, la probabilité π_0 de l'événement vérifie :

$$e^{\beta_0} = \frac{\pi_0}{1+\pi_0}$$

Donc l'exponentielle du terme constant correspond à la cote d'un individu pour lequel $x_i = 0$ Dans de nombreux cas, cela n'a pas de signification car un tel individu ne peut pas exister (par exemple si l'une des variable est l'âge, l'autre la puissance, il s'agirait d'un individu d'âge et de véhicule de puissance nuls). Sauf toutefois si on a pris la précaution de centrer les variables quantitatives, alors cette valeur e^{β_0} correspond à la cote de l'individu moyen (et présentant les modalités codées 0 s'il y a aussi des inputs qualitatifs).

6.4.3.3 Interprétation des odds

Le terme d'intercept s'interprète comme un odds, et les coefficients de régression comme des odds-ratio : lorsque X_j augmente de $d = 1$ unité, l'odds de Y=1 augmente de $exp(\beta_j d)$ (de manière équivalente, le log-odds augmente de$\beta_j d$)).
Dans notre cas :

- L'odds ratio $exp(-0.004897) = 0.995115$ associé à une augmentation d'âge d'un an sur le risque de survenue du premier sinistre, c'est-à-dire une dimunition du risque de $(1-0.995115)*100 = 0.48\%$ pour une variation d'un an d'AGEC.
- L'odds ratio $exp(0.009594) = 1.009640$ associé à une augmentation d'âge d'un an sur le risque de survenue du premier sinistre, c'est-à-dire une augmentation du risque de 0.96% pour une variation d'un an d'AGEV.

6.4.3.4 Résultat :

Les analyses effectuées jusqu'ici nous ont permis d'adopter le ***Modèle 2*** *comme modèle final. Il s'interprète comme suit :*
La probabilité d'avoir une durée de survie inférieure à $\tau = 50$ *jours connaissant l'âge du souscripteur et l'âge et la puissance du véhicule assuré peut être estimée par :*

$$P(Y = 1 \mid AGEC, AGEV, PUISVEH) = \frac{e^{\theta}}{1+e^{\theta}}$$

Avec

$$\begin{aligned}\theta &= \widehat{\beta_0} + \widehat{\beta_1} \quad AGEC + \widehat{\beta_2} \quad AGEV + \widehat{\beta_3} \quad PUISVEH \\ &= -0.214216 - 0.004897 \quad AGEC + 0.009594 \quad AGEV - 0.052367 \quad PUISVEH\end{aligned}$$

6.4.4 Validation du modèle

Maintenant que nous avons retenu un modèle, il reste à le valider ; c'est à dire à mesurer son ajustement à notre base de données et sa capacité de prédiction pour les nouveaux clients.

6.4.4.1 Test d'adéquation par la déviance

En ce qui concerne le contrôle de la légitimité du modèle, nous devons nous intéresser à la déviance standardisée et la comparer au nombre de degrés de liberté des résidus. Il est cependant nécessaire de la standardiser en divisant par l'estimation du paramètre de dispersion qui vaut 1 dans cet estimation. Nous obtenons ici que la déviance standardisée vaut 4353.7 ce qui est n'est pas loin du degrés de liberté (le rapport vaut 1.28 proche de 1). Nous pouvons donc admettre que le modèle est pertinent (on rappelle que le modèle est acceptable si le rapport de la déviance standardisée sur le degré de liberté n'est pas grand devant 1).

Remarque 6.4 La validité de la loi et donc du test n'est qu'asymptotique, il est donc nécessaire d'avoir un peu de recul quant aux conclusions. Lorsque les données sont binaires et qu'aucune répétition n'est présente au point $X = x_i, \forall i$, alors D ne suit pas une loi du χ^2. Pour les données binaires le test d'adéquation d'Hosmer Lemeshow est à conseiller.

6.4.4.2 Test de Hosmer-Lemeshow

Sous R, nous avons écrit une fonction qui calcule la statistique de Hosmer et Lemeshow et la p-value associée, basés sur le vecteur des observations et les probabilités prédites. Nous avons obtenu les résultats suivants :
$\widehat{C} = 3.387192$ et $p - value = 0.9077662$
La p-value est supérieure au risque usuel de 5%. Le modèle est validé, il est compatible avec les données.

6.4.5 Evaluation du pouvoir prédictif du modèle

Nous avons jusque là vu que la régression logistique permet d'estimer la probabilité d'avoir "$Y = 1$" i.e. la probabilité d'avoir une durée de survie inférieure à 50 jours quand on connaît AGEC,SEXE du souscripteur et AGEV,PUISVEH du véhicule assuré.
Sur la base de ces probabilités, on peut définir une règle de classification de la manière suivante :

- Si la probabilité est supérieure à un seuil S_0 fixé, on classe le client comme client "haut sinistré" $(Y = 1)$.
- Si au contraire la probabilité est inférieure ou égale à S_0, le client n'est pas classé comme "haut sinistré" $(Y = 0)$.

Bien que le seuil $S_0 = 0.5$ paraisse à priori une valeur raisonnable, il n'est pas du tout évident que ce soit exact. Pour chaque valeur de S_0 on peut calculer la sensibilité et la spécificité du modèle. La sensibilité est définie comme la probabilité de classer l'individu dans la catégorie $Y = 1$ étant donné qu'il est effectivement observé dans celle-ci :

$$Sensibilité = Pr("haut\ sinistré\ " \mid Y = 1)$$

La spécificité quant à elle est la probabilité de classer l'individu dans la catégorie $Y = 0$ étant donné qu'il est effectivement observé dans celle-ci :

$$Spécificité = Pr("non\ haut\ sinistré" \mid Y = 0)$$

La qualité de la méthode de classification est généralement mesurée par ces deux indicateurs (sensibilité et spécificité) au moyen de la courbe ROC (Receiver Operating Characteristic curve) qui est la courbe représentative de la sensibilité en fonction de (1-spécificité). Ainsi, l'aire au dessous de la courbe ROC nous permet de mesurer globalement la capacité du modèle à affecter correctement les sujets à leurs classes respectives. Le graphe ci-dessous donne la courbe ROC correspondante au modèle retenu **Modèle 2** .

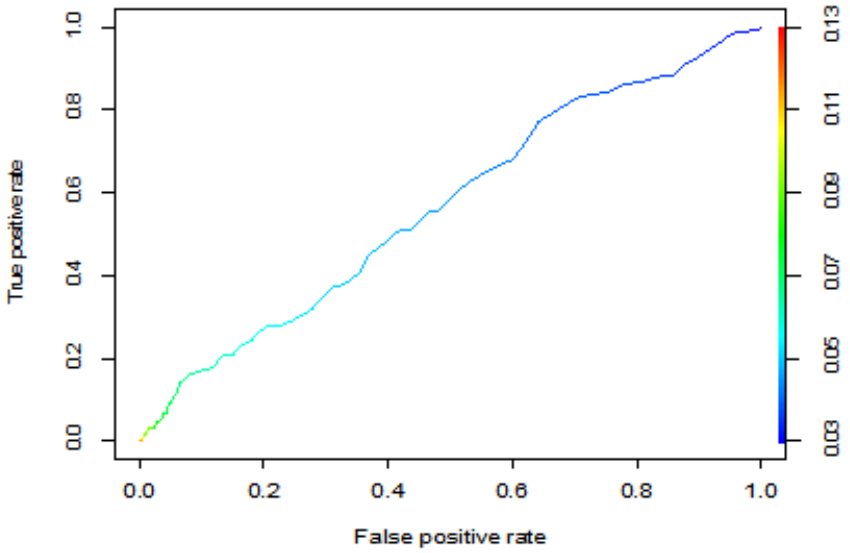

FIGURE 6.6 – Courbe ROC

Ce graphique représente une aire au dessous de la courbe ROC $C_o = 0.687$.
Rappelons que ce coefficient C_0 représentant l'aire au dessous de la courbe ROC pour notre modèle n'est pas loin du critère $0.7 \leq C < 0.8$ correspondant à une discrimination acceptable.
L'aire au dessous de la courbe ROC nous montre que la discrimination n'est pas très bonne ; ceci pouvant s'expliquer par l'existence d'un groupe important d'individus ayant des statuts diffèrents, mais de profils semblables.

Application 2 :
Les modèles de survie

Les modèles de survie, ont été développés pour des applications en biologie, en médecine (Biostatistique, épidémiologie . . .), en démographie (espérance de vie aux divers âges . . .), en économie (analyse du marché de travail, durées de vie des entreprises . . .), en finance (défaillances de crédit), en fiabilité (durée de vie de composants industriels) [11],[17]. Le domaine d'application de ces modèles à l'actuariat de l'assurance est non négligeable. On trouve surtout des applications aux problèmes de durée de vie humaine et la construction de tables d'expérience [20].
Dans cette partie on s'intéressera tout particulièrement à l'étude du phénomène de sinistralité automobile et à l'élaboration d'un modèle prédictif de durée de survie contre la sinistralité. Une application pratique sera réalisée sur un extrait du portefeuille d'une compagnie de taille significative sur le marché Algérien d'assurance non-vie. On entend par durée de survie Auto, la durée séparant la date de survenue du premier sinistre de la date de création de contrat. La date de création de contrat est connue, par contre la date de survenue des sinistres n'est pas connue au moment des traitements pour tous les contrats, on dit que les données sont censurées à droite. Ces données manquantes compliquent sérieusement l'analyse et nous poussent à utiliser des outils plus adéquats : les modèles de survie. Les données utilisées sont présentées dans la section précédente ref.

Mécanisme de censure

L'application des modèles de survie étant tributaire de la caractéristique de censure et vue la structure de notre base de données, nous avons élaboré le mécanisme de censure suivant :
La variable d'intérêt considérée est la durée de survie contre la sinistralité auto (notée DSURV), c'est-à-dire la durée de survenue du premier sinistre, définissant la différence entre la date du premier sinistre et la date de création du contrat telle que :

- Si la durée de survie DSURV est supérieure à la durée de couverture DCOUV de souscripteur qui est définie par la différence entre la date de résiliation et la date de création du contrat ; il s'agit d'une censure à droite fixe.
- La durée de couverture du contrat est indépendante de la survenue de la variable d'intérêt ; on dit qu'il s'agit de censure non informative.

Les méthodes d'inférences statistiques classiques ne sont plus appropriées à l'ensemble de données considérées, en raison de la présence des censures. Afin d'estimer les lois décrivant la durée de survie Auto, particulièrement, la fonction de survie exprimant la

probabilité de survie auto, et éventuellement en présence des quatre variables exogènes : l'âge du véhicule AGEV, l'âge de conducteur AGEC (sous l'hypothèse que le conducteur est le propriétaire du véhicule), le sexe du conducteur SEXE et la puissance du véhicule PUISVEH. On propose d'employer le modèle à hasard proportionnel de Cox.
Pour des raisons d'effectifs, à cette étape il vaut mieux travailler sur des variables qualitatives qu'avec les variables quantitatives disponibles. Pour chaque contrat étudié, nous disposons de différentes variables exogènes :
2. L'âge du souscripteur

$$\text{AGEC=}\begin{cases}1\text{,si l'âge du souscripteur est inférieur à 36 ans,}\\ 2\text{,si l'âge du souscripteur est entre 36 et 50 ans,}\\ 3\text{,si l'âge du souscripteur est supérieur à 50 ans.}\end{cases}$$

3.Sexe du souscripteur

$$\text{SEXE=}\begin{cases}1\text{,si l'assuré est un homme,}\\ 2\text{,si l'assuré est une femme.}\end{cases}$$

4. L'âge du véhicule

$$\text{AGEV=}\begin{cases}1\text{,si l'âge du véhicule est inférieur à 3 ans,}\\ 2\text{,si l'âge du véhicule est entre 4 et 9 ans,}\\ 3\text{,si l'âge du véhicule est supérieur à 10 ans.}\end{cases}$$

5. La puissance du véhicule

$$\text{PUISVEH=}\begin{cases}1\text{,si le véhicule a une petite puissance ,}\\ 2\text{,si le véhicule a une puissance moyenne,}\\ 3\text{,si le véhicule a une grande puissance .}\end{cases}$$

6.5 Approche semi-paramétrique : le modèle de Cox

Désignons par T la variable d'intérêt ($T = DSURV$). Supposons qu'on observe la durée de survenue du premier sinistre de n contrats autos $t_1, t_2, , t_n$ et que δ_i est l'indicatrice de l'occurrence de l'événement d'intérêt, qui prend zéro si la iéme durée de survie t_i, $i = 1, 2, , n$, est censurée à droite et une unité sinon.
La fonction de vraisemblance partielle de Cox est :

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right)^{\delta_i}$$

Où $R(t_i)$ est l'ensemble des contrats à risque à l'instant t_i, x_i le vecteur des quatre variables exogènes au iéme contrat et β est le vecteur des paramètres à estimer, exprimant l'effet de ces variables sur la durée de survie DSURV.

Dans ce cas le rapport $\frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)}$ exprime la probabilité conditionnelle que l'éme souscripteur a eu le premier sinistre en t_i, sachant qu'il y a eu un seul sinistre dans ce temps. La fonction de log-vraisemblance correspondante est alors donnée par

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \left(\beta' x_i) - \log\{\exp(\beta' x_l)\}\right)$$

Le vecteur β des paramètres de ce modèle est estimés par la méthode itérative de Newton-Raphson.

6.5.1 Estimation des paramètres

Nous commençons par illustrer l'estimation d'un premier modèle dans lequel nous cherchons à expliquer la durée DSURV en fonction des facteurs : AGEV, AGEC, SEXE, PUISVEH. L'ajustement d'un modèle de régression à hasards proportionnels sur les données (i.e. l'estimation des paramètres) se fait au moyen de la fonction `coxph` du package `survival`, sous le logiciel R. Les résultats obtenus sont présentés dans le tableau suivant : Le modèle appris sur les données, est

Variable Exogène	$\widehat{\beta}$(coef)	exp(coef)	se($\widehat{\beta}$)	z value	pr(>|z|)
AGEC	-0.21580	0.80590	0.07625	-2.830	0.004656 **
SEXE	0.80290	2.23201	0.38501	2.085	0.037032 *
AGEV	0.23844	1.26926	0.07116	3.351	0.000806 ***
PUISVEH	0.04996	1.05122	0.09395	0.532	0.594917

Rsquare	= 0.335	(max possible	= 0.633)
Likelihood ratio test	= 28.23	on 4 df,	$p = 1.122e-05 \simeq 0$
Wald test	= 26.34	on 4 df,	$p = 2.704e-05 \simeq 0$
Score (logrank) test	= 26.93	on 4 df,	$p = 2.057e-05 \simeq 0$

TABLE 6.6 – Estimation des coefficient du modèle de Cox

6.5.2 Interprétation des coefficients estimés

Les coefficients estimés, leurs erreurs standards, les statistiques du hasard ratio et les p-values associées sont données dans le Tableau 6.6.

- Sous $\widehat{\beta}$**(coef)** on y lit le *coefficient estimé* de chaque facteur explicatif. Celui-ci mesure l'effet du facteur sur le logarithme du risque.
  Les coefficients des deux variables portant sur le souscripteur (AGEC et SEXE) sont négatifs ce qui nous indique que les conducteurs les plus jeunes de sexe masculin ont un niveau bas du risque de sinistralité.
  On peut interpréter ces deux coefficient autrement, par exemple, le coefficient d'AGEC vaut -0.271513, ce qui signifie que le logarithme de risque diminue de -0.271513 en fonction de l'âge du conducteur et de -0.985335 pour les femmes.
  Ainsi plus lâge du véhicule est grand, plus le risque de sinistralité augmente

Covariable	Ratio	Wald $\chi^2_1(1ddl) = 3.84$	Degré de signification	Décision
AGEC	-2.8301	8.0094	0.0046	Effet significatif
SEXE	2.0854	4.3488	0.0361	Effet significatif
AGEV	3.3507	11.2271	0.0008	Effet significatif
PUISVEH	0.5317	0.2827	0.5949	Effet non significatif

TABLE 6.7 – Test de Wald

- La colonne **se($\widehat{\beta}$)** (*Standard Error*) donne l'erreur standard du coefficient qui mesure la variabilité de l'estimateur utilisé.
- Les deux colonnes **z value** et **pr(>|z|)** concernent la statistique du ratio critique utilisé pour tester la significativité individuelle de chaque coefficient. Elles donnent respectivement la valeur de la statistique et la p-value associée.

6.5.3 Evaluation du modèle

La pertinence statistique d'un modèle se fonde en règle générale sur la significativité statistique individuelle des coefficients, sur l'ajustement global et sur l'analyse des résidus.

6.5.3.1 Significativité individuelle des coefficients

La procédure de régression de Cox donne pour chaque coefficient le ratio critique, souvent appelé t de Student (statistiue de student), entre le coefficient et son erreur standard que l'on examine pour juger sa significativité statistique. Une règle sommaire avec le ratio critique est de considérer le coefficient comme significatif lorsque le ratio est supérieur à 2.
Pour un coefficient individuel, la statistique de Wald est simplement le carré du ratio critique. Par exemple le coefficient de SEXE est 0.80290 et son erreur standard 0.38501 , Le ratio critique vaut $0.80290/0.38501 = 2.0854$ dont le carré est 4.3488 soit la valeur indiquée pour la statistique de Wald est ici grande puisque supérieure $2^2 = 4$. Elle indique donc un effet significatif du genre (sexe). Pour une règle plus précise, on compare la valeur de Wald pour un coefficient à un chi-deux à 1 degré de liberté. En effet, sous l'hypothèse d'effet nul, la statistique de Wald d'un coefficient est distribuée comme un khi-deux à 1 degré de liberté. La significativité de la variable SEXE est ainsi confirmée par un degré de signification (Inférieure à $\alpha = 0.05$) :

$$sig(SEXE) = P(> 4.348) = 0.0361$$

On résume pour chaque facteur de régression les valeurs du ratio et la statistique de Wald calculée selon la règle précédente dans le tableau suivant :

$$\text{pour } i = 1,..,p : \qquad H_0 : \beta_i = 0 \qquad vs \qquad H_1 : \beta_i \neq 0$$

D'après les résultats du Tableau 6.7 on conclut que les effets des covariables AGEC,SEXE et AGEV sont significatifs, c'est-à-dire que le hasard risque s'est influencé.

6.5.3.2 Evaluation globale

Lorsque les paramètres β sont estimés, trois tests, asymptotiquement équivalents, permettent de déterminer si les coefficients β estimés sont significativement différents de 0 (test de l'hypothèse nulle des β). Il s'agit du test de Wald (maximum de vraisemblance), le test du rapport de vraisemblance et test du score [Therneau et Grambsch 2000].

i. Les statistiques de Khi-deux

Les indications sur l'ajustement global du modèle sont fournies par la fonction de régression sous R .On y trouve :
- La statistique du khi-deux du rapport de vraisemblance (Likelihood ratio test).
- La statistique du khi-deux du test de Wald.
- La statistique du khi-deux du test du Score.

Ces statistiques permettent d'évaluer si globalement l'ensemble des facteurs explicatifs considérés améliore significativement l'ajustement du modèle naïf qui ne tient compte d'aucun facteur. En d'autres termes, pour un modèle avec p coefficients, ils permettent de tester l'hypothèse :
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad vs \qquad H_1$:qu'un au moins des coefficients est non nul.
Sous l'hypothèse H_0 (modèle naïf correct), les trois statistiques sont distribuées asymptotiquement selon une loi du khi-deux à p degrés de liberté. On considère donc l'amélioration par rapport au modèle naïf comme significative lorsque la valeur de ces statistiques est suffisamment grande, soit lorsque leur degré de signification est inferieur, en règle générale, à 5%.
Le degré de signification est ici défini comme la probabilité que le khi-deux prenne une valeur supérieure Ã la valeur observée de la statistique. Dans notre cas la valeur du Score est 26.93 supérieure à $\chi^2_4 = 9.48$, ce que confirme le degré de signification :

$$sig(Score) = P(\chi^2_4 > Score_{obs}) = P(\chi^2_4 > 26.93) = 2.053729.10^{-05}$$

On remarque que les valeurs des statistiques de Wald et rapport de vraisemblance sont aussi supérieures à χ^2_4, et ceci confirmées par les probabilités presque nulles. Le modèle naïf doit donc être rejeté au profit du modèle ajusté. Cela ne signifie pas que le modèle ajusté est satisfaisant, mais nous dit simplement que le modèle ajusté fait mieux que le modèle naïf.

ii. Pseudo R^2

Il s'interprète plus ou moins comme la proportion de réduction du défaut d'ajustement ou "dispersion résiduelle" du modèle naïf. Dans notre cas, on a $R^2 = 0.335$. La part "expliquée" de la "dispersion" totale est de l'ordre de (33.5%).

6.5.4 Test de l'hypothèse de proportionnalité

Le modèle semi paramétrique de Cox est très général puisqu'il ne suppose aucune hypothèse sur la distribution des durées des épisodes. Il suppose cependant la proportionnalité des risques. Il convient donc de vérifier que cette hypothèse est

raisonnable. Une première approche consiste à vérifier la proportionnalité graphiquement. Une seconde repose sur des tests statistiques. Les tests graphiques examinent les effets, c'est-à-dire les covariables introduites dans le modèle, individuellement. Les tests statistiques portent également sur les effets individuels.

6.5.4.1 Examen graphique des résidus partiels

L'analyse graphique des résidus partiels de Schönefeld constitue une alternative mieux à même de mettre en évidence les situations de non-proportionnalité des risques. L'idée est qu'en cas de proportionnalité des risques, l'écart entre le profil d'un cas i et le profil moyen des cas exposés en t_i devrait être aléatoire et indépendant de t_i. On ne devrait donc pas observer d'effets systématiques dans l'évolution des résidus partiels avec la durée. Pour ce fait on examine le diagramme de dispersion des résidus partiels selon la durée t, augmenté de la droite de régression des résidus sur t.
Nous avons généré les graphiques des résidus pour les variables significatives du modèle du Cox estimé AGEC et AGEV.

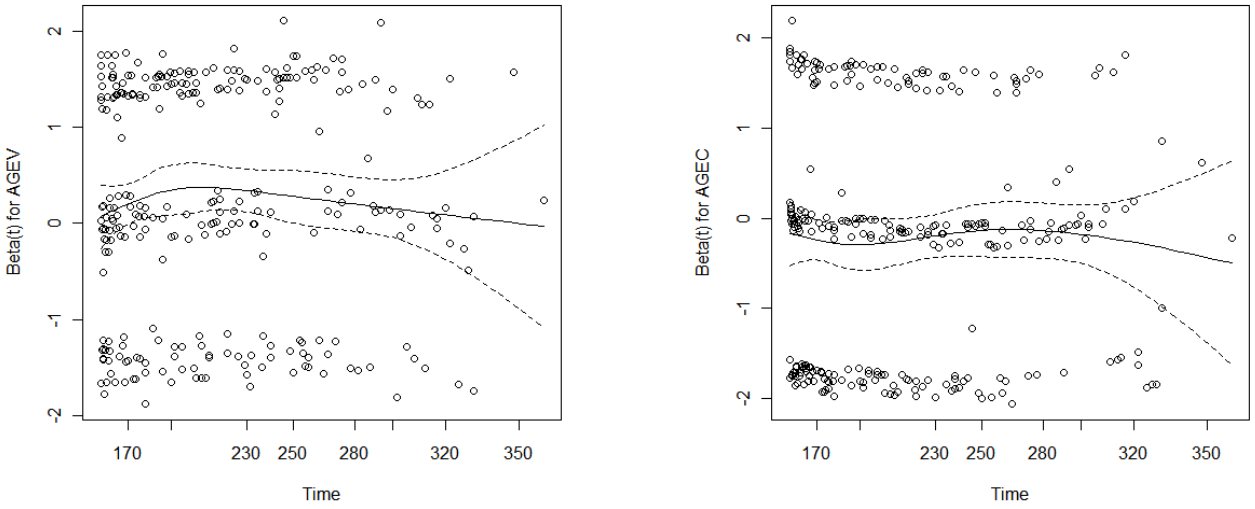

FIGURE 6.7 – Résidus partiels pour AGEV et AGEC

Ces graphiques décrivent l'évolution des résidus en fonction du temps (jours). La droite de régression indique la tendance. Les pentes sont légèrement positives, indiquent que les résidus sont positifs pour les conducteurs âgés et pour les anciens véhicules.

6.5.4.2 Test statistique

Il consiste à tester si la pente de la droite de régression des résidus partiels sur la durée t est statistiquement significative.
Nous avons obtenu les résultats suivants :
Les p-values sont assez grand pour toutes les variables ce qui nous amène à admettre l'hypothèse de proportionnalité des risques.

Covariable	Rho	chisq	p-value
AGEC	0.00807	0.019097	0.890
SEXE	-0.04856	0.704298	0.401
AGEV	-0.00124	0.000433	0.983
PUISVEH	0.01352	0.048837	0.825

TABLE 6.8 – Test de proportionalité

6.5.5 Estimation de la fonction de survie

Le modèle de Cox donne l'estimation $\widehat{\beta}$ des coefficients β. Pour obtenir une estimation de la probabilité de survie $S(t,z)$, il nous faut encore une estimation de la fonction $S_0(t)$ de référence. En l'absence d'hypothèses sur la forme de la distribution, on estime $S_0(t)$ de façon non paramétrique, avec un estimateur de Kaplan-Meier.

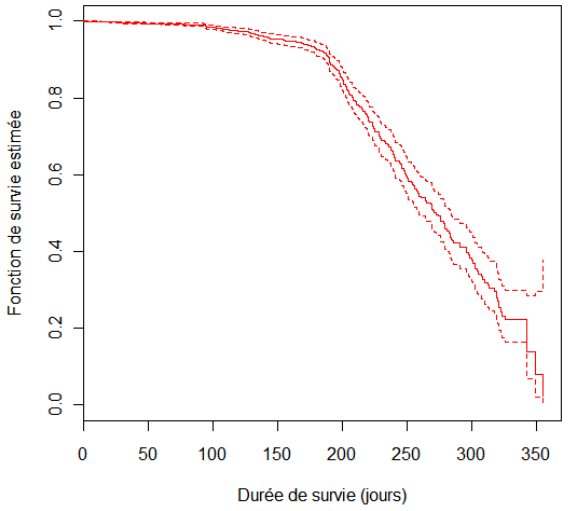

FIGURE 6.8 – Estimation de la fonction de survie correspondant au modèle de Cox

6.6 Conclusion

Ce chapitre nous a permis dans un premier temps de déterminer un groupe de variables pouvant expliquer de manière significative (au seuil 10 %) la sinistralité. L'analyse simultanée de ce groupe de variables par le modèle logistique ressort les plus pertinentes et plus discriminantes pouvant représenter valablement le groupe : l'âge du souscripteur et l'âge et la puissance du véhicule. Après examen des résultats de l'analyse, on peut noter que :

- Les contrats des véhicules d'une puissance minimale sont les plus responsable des sinistres produits dans la période $[0, 50[$ *jours* de l'inscription à la compagnie d'assurance ;
- L'âge du véhicule et l'âge du souscripteur sont associés à la sinistralité ; les anciens véhicules et les jeunes conducteurs étant les plus exposés au risque.

En appliquant le modèle de survie de Cox sur notre base de données, éventuellement en présence de variables exogènes enregistrées pour chaque contrat Auto, on constate que certaines variables ont un impact temporel sur la variable d'intérêt, et elles nous ont permet de préciser l'évolution de la durée de survie contre la sinistralité.
En l'absence d'information à priori sur la forme de la fonction de survie, nous l'avons estimée par la méthode non-paramétrique de Kaplan-Meier.

Conclusion générale

Ce travail a pour objet la présentation des principales techniques statistiques utilisées pour l'analyse des durées de réalisation d'un ou de plusieurs événements d'intérêt. Nous avons étudié deux approches : les modèles statistiques usuels et les modèles de régression.
Nous avons entrevu à travers ce document une méthodologie de l'analyse de la segmentation en assurance automobile. Nous avons tout d'abord compris que la maitrise de la segmentation est primordiale pour préserver son portefeuille ou conquérir de nouveaux assurés au sein de la première source de chiffre d'affaires en assurances de biens et de responsabilité. Nous avons également constaté que dans un univers d'innovation constante, la segmentation des risques semble devenir de plus en plus poussée.
A partir de ces constats, nous avons cherché à mettre en avant les différentes étapes de la segmentation du risque automobile à travers les modèles de régression logistiques. En effet l'introduction de ces modèles nous a permis d'évaluer l'impact des variables exogènes sur la durée de survie d'un contrat Auto.
Un modèle semi-paramétrique de Cox a été utilisé sur différentes variables exogènes lorsque les hypothèses liées au modèle le permettaient. L'impact temporel de certaines variables nous a permis en utilisant le modèle de Cox, de préciser l'évolution de la durée de vie de ces contrats. Enfin en l'absence d'information a priori sur la forme de la fonction de survie, nous l'avons estimée par la méthode non-paramétrique de Kaplan-Meier.
En conclusion, cette étude sur le phénomène de survenue des sinistres nous a permis d'illustrer les méthodes classiques d'estimation des durées de vie appliquées à un portefeuille d'une compagnie d'assurance non-vie.

Bibliographie

[1] AALEN O.O.[1975], *Statistical Inference for a Family of Counting Processes*. PhD thesis, University of California, Berkeley.

[2] AALEN O.O., BORGAN O.,GJESSING H.K. [2008], *Survival and Event History Analysis*, Springer-Verlag, New York.

[3] AGERSTI A.[2002], *An Introduction to Categorial Data Analysis*. New York, Wiley, p.710.

[4] AKAIKE H.[1973], *Information Theory and an Extension of the Maximum Likelihood Principle*
In Proceedings of International Symposium on Information Theory. B.N. Petrov et F. Czaki,Budapest.

[5] ANDERSON P.K.,BORGAN O.,GILL R.D.,KEIDING N.[1991], *Statistical Models Based on Counting Processes*. Springer-Verlag.

[6] BOUKHETALA K., GUIDOM A. [2012], ***Sim.DiffProc*** *: Simulation of Diffusion Processes*. R package version 2.5, http://CRAN.R-project.org/package=Sim.DiffProc.

[7] BOUKHETALA K.[2011], *Processus Aléatoires Appliqués à la Finance et l'Actuariat*. Cours de Post-Graduation de Probabilités et Statistiques, Université des Sciences and Technologie Houari Boumedienne (USTHB)

[8] BRESLOW N.[1974], *Covariance Analysis of Censored Survival Data*. Biometrics, vol. 30 p. 89-99.

[9] CAIRNS A., BLAKE D., DOWD K., [2004], *Pricing Frameworks for Securitization of Mortality Risk*. AFIR.

[10] COX,D.R.[1972], *Regression Models and Life-Tables*. Avec discussion, Journal of the Royal Statical Society, Series B,vol.74 p.187-220 .

[11] DROESBEKE J.J., FICHET B., TASSI P. [1989], *Analyse Statistique des Durées de Vie : Modélisation et Données Censurées*. Economica, ISBN 13 : 978-1-4020-5953-7.

[12] DUPUY J.F.[2002], *Modélisation Conjointe de Données Longitudinales et de Durées de Vie*. Université de Paris V, Thèse de doctorat.

[13] DUYME F., CLAUSTRIAUX J.J.[2003], *La régression Logistique Binaire*. Notes Stat.Inform (Gembloux), 2004/6, p.26.

[14] FLEMING T.R., HARRINGTON D.P.[1991], *Counting Processes and Survival Analysis*. Wiley Series in Probability and Mathematical Statistics, New-York : Wiley.

[15] HOSEMER D.W., LEMSHOW S.[2000], *Applied Logistic Regression*. New-York , Wiley, p.392.

[16] HUBER C. *Cours de Modélisation Biostatique en S-plus*. Université de Paris, René descartes UFR Biomédical, ISBN 978-3-642-20310-7.

[17] KAPLAN E.L, MEIER P [1958], *Non-Parametric Estimation from Incomplete Observations*. Journal of the American Statistical Association, vol. 53,p. 457-481.

[18] LEE R.D, CARTER L.[1992], *Modelling and Forecasting the Time Series of US Mortality*. Journal of the American Statistical Association, 87,659-671.

[19] NELSON W. B. [1972], *Theory and Applications of Hazard Plotting for Censored Data*. Technometrics, vol. 14 p. 945-965.

[20] PLANCHET F, THEROND P [2006], *Modéles de Durée. Application Actuarielles* . Economica, ISBN 0-387-54062-8.

[21] SCHORNFELD D [1982], *Partial Residuals for the Proportional Hazards Regression Model*. Biometrika, vol. 69 p. 239-241.

[22] SHORACK, G.R. J.A.WELLNER [1986], *Empirical Processes with Applications to Statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics, New York : John Wiley Sons Inc.

[23] THERNEAU T.M., GRAMBSCH P.M., FLEMING T.R[1990], *Martingale-Based Residuals for Survival Models*. Biometrika, vol. 77, $n^0$1, p. 147-160.

[24] THERNEAU T. M., GRAMBSCH P. M.[2000], *Modeling Survival Data : Extending the Cox Model*. Series : Statistics for Biology and Health, New-York : Springer.

[25] VAART, A.W.VAN DER J.A. WELLNER [1996], *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics : New York : Springer-Verlag.

[26] VAART, A.W.VAN DER[1998], *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge : Cambridge University Press.