Aternatives au modèle de Cox
Application en assurance automobile

Farid Beninel*, Jean-Marie Marion**

*LMA Futuroscope - UMR CNRS 7348
BP 30179, 86962 Futuroscope-Chasseneuil Cedex
fbeninel@gmail.com

**IMA-UCO, rue Rabelais, 49003 Angers
jean-marie.marion@uco.fr

Résumé. Ce travail traite de méthodologie, pour l'étude de durée de vie de contrats d'assurance VAM. L'approche développée est celle de modèles de durée, selon les profils de description. Les données consistent en une grande base, correspondant au portefeuille automobile d'une grande compagnie d'assurance européenne. On teste, par différentes approches, l'hypothèse de proportionnalité des hasards instantanés, à la base du modèle de Cox s'avérant inadéquat. On montre que pour ces données, le choix se porte sur un modèle faisant évoluer les paramètres en fonction du temps : le modèle de Aalen.

1 Introduction

Dans ce travail, on sintéresse au phénomène de résiliation de contrat d'assurance automobile et en particulier, aux tests de modèles prédictifs de durée de vie de contrats. On entend par durée de vie de contrat `Auto`, la durée séparant la date de résiliation de la date de création du contrat. La date de création de contrat est toujours connue, tandis que la date de résiliation n'est pas toujours connue *i.e.,* les données sont censurées à droite. Ces données censurées, non assimilables à des données manquantes, conduisent à utiliser des outils spécifiques : les modèles de durée (*cf.* Therneau et Grambsch (2001), Planchet et Thérond (2006) et Klein et Moeschberger (2003)).
Les modèles de durée sont utilisables dès lors qu'il s'agit de modéliser le temps qui s'écoule entre deux événements à partir d'observations de durée et éventuellement de variables explicatives dites variables exogènes ou covariables. L'analyse des durées de vie peut se faire en utilisant des méthodes d'inférence statistique traditionnelles adaptées aux données censurées ((Planchet et Thérond, 2006),(Klein et Moeschberger, 2003)) (modèles paramétriques, semi-paramétriques et non-paramétriques). Ces méthodes considèrent que l'on observe des variables aléatoires positives représentant des durées jusqu'à ce qu'un certain évènement ait lieu (ici, la résiliation du contrat).
Il est aussi possible d'aborder les modèles de survie sous forme de processus ponctuels, en considérant les observations comme des processus évoluant au cours du temps. On peut, par exemple associer, à un contrat, un processus « de présence à risque » qui vaut 1, à chaque

instant où le contrat est observé et 0, s'il n'a pas encore été résilié. Aussi, on peut associer un processus qui vaut 1, seulement à partir de l'instant où le contrat est résilié 0 sinon. Ces méthodes permettent l'usage de résultats puissants concernant les martingales et les processus prévisibles pour procéder à des estimations et des tests sur les durées de vie. Pour plus de détails, on peut se référer par exemple à Klein et Moeschberger (2003) ou à Andersen et al. (1993).
Les modèles de survie ont été développés pour des applications en biologie, en médecine et épidémiologie, en démographie (espérance de vie aux divers âges,...), en économie (analyse du marché de travail, durées de vie des entreprises, ...), en finance (durée jusqu'à défaut de paiement,...), en assurance vie et prévoyance (tables de mortalité, ...), en fiabilité (durée de vie de composants industriels, ...) et autres domaines (Therneau et Grambsch (2001), Planchet et Thérond (2006) et Klein et Moeschberger (2003)).
Parmi les différents modèles de durée de vie, utilisés en actuariat de l'assurance, figurent les modèles à hasard proportionnel, en particulier le modèle de Cox. Il s'agit de modèles semi-paramétriques dans lesquels on modélise la fonction de hasard, en considérant que les covariables agissent par effet multiplicatif *via* une fonction monotone à valeurs positives, sur une fonction de hasard de base. Losque cette hypothèse de proportionnalité des hasards n'est pas vérifiée, on considère des alternatives comme la partition de modèles ou modèles stratifiés, les modèles à coefficients dépendant du temps ou encore le modèle additif de Aalen.

Cet article est organisé comme suit : En section 2, on présente le modèle de Cox, les différents tests de validité associés, ainsi que différents modèles alternatifs. En en section 3, on présente le modèle de Aalen.La section 4 est dévolue à l'application, à partir d'une base de données réelles où les individus statistiques sont des contrats et la description est relative au contrat et à son souscripteur. L'expérimentation est orientée tests de validité du modèle à hasard proportionnel (modèle de Cox) et illustration d'une meilleure adéquation du modèle de Aalen.

2 Modèles à hasard proportionnel

2.1 Le modèle théorique

Considérons le modèle où la fonction de hasard α est telle que

$$\alpha(t;\mathbf{Z}) = \alpha_0(t)\Psi(\mathbf{Z};\beta),$$

avec
-$\mathbf{Z}$ vecteur *p*-dimensionnel des covariables, indépendant du temps,
-β vecteur des paramètres associés,
-α_0 fonction de hasard de base, indépendante des covariables,
-$\Psi(\mathbb{R}^p \mapsto \mathbb{R}^+)$ fonction de lien, traduisant l'effet multiplicatif des covariables sur le risque instantané (ou hasard) de base. La fonction Ψ la plus utilisée est celle de Cox (1972) :

$$\Psi(\mathbf{Z};\beta) = \exp(\beta^T\mathbf{Z}).$$

Pour ce modèle, le rapport des fonctions de hasard ou *risk ratio*, pour deux individus décrits par les vecteurs de covariables, *resp.* $\mathbf{Z}_1$ et $\mathbf{Z}_2$, donné par

$$\frac{\alpha(t;\mathbf{Z}_1)}{\alpha(t;\mathbf{Z}_2)} = \frac{\Psi(\mathbf{Z}_1;\beta)}{\Psi(\mathbf{Z}_2;\beta)} \qquad (\mathcal{PH}),$$

est indépendant du temps.
Remarquons que, pour le modèle de Cox, ce rapport est égal à $exp(\beta^T(\mathbf{Z}_1 - \mathbf{Z}_2))$. La relation $(\mathcal{PH})$ traduit l'hypothèse du hasard proportionnel.

2.2 Estimation des paramètres du modèle de COX

Soit l'échantillon de triplets $(T_j, \delta_j, \mathbf{Z}_j)_{1\leq j\leq n}$ où $T_j = \inf(X_j, C_j)$ avec
$-C_j$: la censure droite et X_j la durée de vie du $j^{ème}$ contrat d'assurance,
$-\delta_j$: l'indicatrice de résiliation,
$-\mathbf{Z}_j$: le vecteur des covariables, pour le $j^{ème}$ contrat de l'échantillon.

Soit
$-t_1, ..., t_n$: l'échantillon des durées, dans lequel on a D résiliations et $n - D$ censures,
$-0 < X_1^* < X_2^* < \ldots < X_D^*$: l'échantillon ordonné des D instants de résiliation, supposés distincts,
$-\mathbf{Z}_j^*$: le profil ou vecteur, réalisation des covariables, associé au contrat résilié en X_j^*,
$-\Re(X_j^*)$: l'ensemble des contrats à risque, en X_j^{*-} *i.e.*, $\Re(X_j^*) = \{i : X_i \geq X_j^*\}$,
$-\mathcal{E}(X_j^*)$: l'évènement représentant la résiliation, au bout d'une durée X_j^*, d'un contrat appartenant à $\Re(X_j^*)$,
$-\mathcal{E}(Z_i^*)$: l'évènement représentant la résiliation du contrat, au profil associé Z_i^* ; ce profil comporte les caractéristiques du contrat et des renseignements sur le souscripteur.

Soit $L_j(\beta)$ la contribution multiplicative, à la vraisemblance, de l'observation $(X_j^*, \mathbf{Z}_j^*)$ *i.e.*,
$$L_j(\beta) = \mathbb{P}(\mathcal{E}(Z_i^*) \mid \mathcal{E}(X_j^*)).$$

S'agissant d'un modèle de COX, on obtient
$$L_j(\beta) = \frac{\exp(\beta^T Z_j^*)}{\sum_{i\in\Re(X_j^*)} \exp(\beta^T \mathbf{Z}_i)}.$$

Par suite, la vraisemblance partielle de COX est donnée par
$$L(\beta) = \prod_{j=1}^{D} L_j(\beta) = \prod_{j=1}^{D} \frac{\exp(\beta^T \mathbf{Z}_j^*)}{\sum_{i\in\Re(X_j^*)} \exp(\beta^T \mathbf{Z}_i)}.$$

On ne perd pas en généralité, en utilisant cette vraisemblance, dévolue au cas où les instants de résiliation sont distincts. Dans le cas de résiliations simultanées, on utilise l'extension de la vraisemblance d'Efron ou celle de Breslow ou encore celle de Cox (Klein et Moeschberger (2003)).
Dans le cas de données censurées à droite, la vraisemblance partielle peut aussi s'écrire
$$L(\beta) = \prod_{j=1}^{n} \left[\frac{\exp(\beta^T \mathbf{Z}_j)}{\sum_{i\in\Re(t_j)} \exp(\beta^T \mathbf{Z}_i)}\right]^{\delta_j}.$$

Présentons la recherche de l'estimateur de β comme se faisant à partir de la log-vraisemblance partielle *i.e.*,
$$\log(L(\beta)) = \sum_{j=1}^{n} \delta_j [\sum_{h=1}^{p} \beta_h Z_{jh} - \log(\sum_{i\in\Re(X_j^*)} \exp(\sum_{h=1}^{p} \beta_h Z_{ih}))].$$

Ainsi, l'estimation de β , par maximum de vraisemblance, sera solution du système aux p équations de vraisemblance

$$\frac{\partial \log L}{\partial \beta_h} = \sum_{j=1}^{n} \delta_j [Z_{jh} - \frac{\sum_{i \in \Re(t_j)} Z_{ih} \exp(\sum_{h=1}^{p} \beta_h Z_{ih})}{\sum_{i \in \Re(t_j)} \exp(\sum_{h=1}^{p} \beta_h Z_{ih})}] = 0, \quad h = 1, \ldots, p.$$

Ce système, aux équations non linéaires, en le vecteur β, se résoud par des procédés itératifs dérivés de la méthode du gradient (algorithme de Newton-Raphson ; algorithme de Nelder-Mead,...).

2.3 Tests de l'hypothèse "hasard proportionnel"

2.3.1 Tests graphiques du *hazard plotting* : représentation log

Etant donné un profil $\mathbf{Z}$, ce test utilise une représentation logarithmique du nuage $\{(t_i, A(t_i; \mathbf{Z}_i)),\ i = 1, \ldots n\}$. Soit $A(.; \mathbf{Z})$ la fonction de hasard cumulé, associée au profil $\mathbf{Z}$; elle est définie par

$$A(t; \mathbf{Z}) = \int_0^t \alpha(s; \mathbf{Z}) ds, \quad t > 0.$$

Dans le cas du modèle de Cox, sa forme explicite est donnée par

$$A(t; \mathbf{Z}) = A_0(t) \exp(\beta^T \mathbf{Z}),$$

avec $A_0(t) = \int_0^t \alpha_0(s) ds$ la fonction de hasard cumulé de base.

Disposant de profils $\mathbf{Z}_i$ différents, on déduit la relation

$$\log(A(t; \mathbf{Z}_i)) - \log(A(t; \mathbf{Z}_j)) = \beta^T (\mathbf{Z}_i - \mathbf{Z}_j).$$

Etant donné β et les profils $\mathbf{Z}_i, \mathbf{Z}_j$, la quantité $\beta^T(\mathbf{Z}_i - \mathbf{Z}_j)$ est indépendante du temps.
Par conséquent, si l'hypothèse ($\mathcal{PH}$) est vérifiée, la représentation graphique des fonctions $\log(A(t; \mathbf{Z}))$ selon le temps t, présentera des courbes de même allure, translatées les unes par rapport aux autres.
En pratique, on utilise des estimateurs de la fonction de hasard cumulée (par exemple, l'estimateur de Nelson-Aalen) et on s'assure que les écarts entre les courbes correspondant aux estimateurs du logarithme des fonctions de hasard cumulé, en fonction du temps, sont à peu près constants.

2.3.2 Tests graphiques du *hazard plotting* : représentation *log-log*

Cette représentation est équivalente à la précédente ; étant donné un profil, on représente le nuage $\{t_i, S^{-1}(t_i; \mathbf{Z_i})\}$ *via* une échelle *log-log*. Du fait que $A(t; \mathbf{Z}) = -\log S(t; \mathbf{Z})$ où S représente la fonction de survie, on peut écrire, dans le cas du modèle de Cox,

$$\log(-\log(S(t; \mathbf{Z}))) = \beta^T \mathbf{Z} + \log(A_0(t)).$$

On en déduit la relation,

$$\log(-\log(S(t; \mathbf{Z}_1)) - \log(-log(S(t; \mathbf{Z}_2))) = \beta^T (\mathbf{Z}_1 - \mathbf{Z}_2).$$

On estimera la fonction de survie, par exemple, par la méthode de Kaplan-Meier pour les différents niveaux de la covariable (ou profils) et on représentera les courbes avec une échelle

log-log en fonction du temps. Sous l'hypothèse ($\mathcal{PH}$), ces représentations consisteront en des courbes translatées, les unes par rapport aux autres.

2.3.3 Tests basés sur les résidus

a) Résidus de Cox-Snell
Le test se base sur le fait que si X est une variable aléatoire de durée et A la fonction de hasard cumulé associée, alors $A(X)$ suit une loi exponentielle de paramètre 1.
On définit les résidus de Cox-Snell, par

$$r_j = \hat{A}_0(T_j) \exp(\hat{\beta}^T Z_j), \quad j = 1, \ldots, n,$$

avec $\hat{\beta}$ estimateur obtenu par maximisation de la vraisemblance partielle de Cox et $\hat{A}_0(T_j)$ un estimateur de la fonction de hasard cumulée de base (par exemple, l'estimateur de Breslow).
Pour que les $r_j (1 \leq j \leq n)$ soient proches d'un échantillon censuré de la loi exponentielle de paramètre 1, on calcule un estimateur de la fonction de hasard cumulé des r_j (par exemple Nelson-Aalen).
Une représentation graphique des $\hat{A}(r_j)$ en fonction des r_j trop éloignée de la première bissectrice conduira à rejeter l'hypothèse ($\mathcal{PH}$).

b) Résidus de Schoenfeld
soit $0 \leq t_{(1)} \leq t_{(2)} \leq ... \leq t_{(n)}$ l'ordre sur les durées et $\mathbf{Z}_{(1)}, \mathbf{Z}_{(2)}, \ldots, \mathbf{Z}_{(n)}$ les profils associés.
Le résidu de Schoenfeld $r_{Sh}(j)$ est un vecteur d'écarts entre composantes de $Z_{(j)}$ et une moyenne pondérée des vecteurs profils Z_i, des individus à risque, en $t_{(j)}$ (le coefficient de pondération est la contribution, de l'observation $(t_i, \mathbf{Z}_i)$, à la vraisemblance maximale) ***i.e.,***

$$r_{Sh}(j) = \delta_j (Z_{(j)} - a_{(j)}),$$

avec $a_{(j)} = \sum_{i \in \Re(t_{(j)})} \dfrac{\exp(\hat{\beta}^T Z_i)}{\sum_{i \in \Re(t_{(j)})} \exp(\hat{\beta}^T Z_i)} Z_i.$

Remarques :
- Les résidus de Schoenfeld sont de somme nulle et sont non corrélés, les uns avec les autres (dans le cas de grands échantillons).
- Si l'hypothèse ($\mathcal{PH}$) est valide, une représentation graphique des résidus $r_{Sh}(j)$, en fonction du temps, ne doit pas faire apparaître de dépendance temporelle.

c) Résidus de Schoenfeld normalisés
Le paramètre β figurant dans le modèle de Cox ne doit pas dépendre du temps pour que l'hypothèse ($\mathcal{PH}$) soit conservée.
Grambsch et Therneau (1994) considèrent $\beta(t) = \beta + \theta * g(t)$ (θ un vecteur *p*-dimensionnel, g une fonction perturbante (*i.e.*, dans le voisinage de zéro), à valeurs dans $\mathbb{R}^p$ et $*$ le produit d'Hadamard).
Pour vérifier la validité de l'hypothèse ($\mathcal{PH}$), un test du score, basé sur un estimateur des moindres carrés généralisés de θ, est utilisé : on teste l'hypothèse $H_0 : \ \beta(t) = \beta$, pouvant être reformulée en $H_0 : \ \theta = 0$.

Les résidus de Schoenfeld normalisés sont définis par $\hat{V}_{(j)}^{-1} r_{Sh}(j)$ où $\hat{V}_{(j)}$ est l'estimation de la matrice des covariances du résidu de Schoenfeld $r_{Sh}(j)$; Grambsch et Therneau (1994) établissent que, $\mathbb{E}(\hat{V}_{(j)}^{-1} r_{Sh}(j)) \simeq g(t_{(j)})$.
Ainsi, l'on peut utiliser la représentation graphique de $\hat{V}_{(j)}^{-1} r_{Sh}(j) + \hat{\beta}$ en fonction de $t_{(j)}$ ou (j), qui révèle la forme fonctionnelle de $\beta(t)$: on rejettera l'hypothèse $(\mathcal{PH})$, lorsque cette représentation s'écarte significativement d'une droite horizontale.

2.4 Les extensions de modèles à hasard proportionnel

a) Modèles stratifiés
Il arrive que l'hypothèse $(\mathcal{PH})$ ne soit vérifiée que par strate (les strates étant définies indépendamment de l'échantillon). Dans le cas de K strates, la fonction de hasard s'écrit
$\alpha_{0k}(t) \exp(\beta^T \mathbf{Z})$, pour $1 \leq k \leq K$, avec α_{0k} fonction de hasard de base associée à la strate k et β paramètre de régression commun à toutes les strates.

La vraisemblance partielle, basée sur un n-échantillon de quadruplets $(T_j, \delta_j, Z_j, S_j)$ où S_j désigne la strate j, $1 \leq j \leq n$, est donnée par

$$L(\beta) = \prod_{k=1}^{K} \prod_{j=1}^{n} \left(\frac{\exp(\beta^T Z_{(j)})}{\sum_{i \in \Re(t_{(j)})} \exp(\beta^T Z_i)}\right)^{\delta_j} 1_{\{S_j = k\}}.$$

b) Modèles avec covariables dépendant du temps
Il est possible d'étendre le modèle de Cox, en supposant que les covariables dépendent du temps ; la fonction de hasard s'écrit

$$\alpha(t; \mathbf{Z}(t)) = \alpha_0(t) \exp(\beta^T \mathbf{Z}(t)), \quad t > 0.$$

L'hypothèse $(\mathcal{PH})$ n'est plus vérifiée, puisque $\dfrac{\alpha(t; \mathbf{Z}_1(t))}{\alpha(t; \mathbf{Z}_2(t))} = \exp(\beta^T(\mathbf{Z}_1(t) - \mathbf{Z}_2(t)))$ dépend de t.
La vraisemblance partielle s'écrit

$$L(\beta) = \prod_{j=1}^{n} \left(\frac{\exp(\beta^T Z_{(j)}(t_{(j)}))}{\sum_{i \in \Re(t_{(j)})} \exp(\beta^T Z_i(t_{(j)}))}\right)^{\delta_j}.$$

c) Modèles avec coefficients dépendant du temps
La fonction de hasard s'écrit

$$\alpha(t; \mathbf{Z}) = \alpha_0(t) \exp(\beta^T(t) Z);$$

l'hypothèse $(\mathcal{PH})$ n'est pas, non plus, vérifiée.

3 Une alternative au modèle à hasard proportionnel : Le modèle de Aalen

3.1 Le modèle théorique

Afin de définir le modèle de Aalen, l'on introduit les processus ponctuels (Andersen et al. (1993)) et plus particulièrement, le processus de comptage $N(t) = (N_i(t))_{1 \leq i \leq n}$ avec $N_i(t) =$

$|\{T_i \leq t : \ \delta_i = 1\}|$, le nombre de contrats résiliés ou censurés, au plus tard en t.
Soit le processus prévisible, associé aux observations, $(1_i(t))_{1\leq i\leq n}$, avec $1_i(t) = 1_{\{T_i \geq t\}}$.
Considérons la filtration $(F_t)_{t\geq 0}$, la fonction de hasard α et le processus intensité, associé à la $i^{ème}$ observation, $\lambda_i(t; \mathbf{Z}_i)$ défini par

$$\lambda_i(t; \mathbf{Z}_i)dt = \mathbb{E}(dN_i(t)/F_{t^-}) = 1_i(t)\alpha(t; \mathbf{Z}_i)dt.$$

Le processus intensité multidimensionnel est le vecteur $(\lambda_i(t; \mathbf{Z}_i))_{1\leq i\leq n}$. Si $\mathbf{Z}_i(t) = (Z_{i1}(t), ..., Z_{ip}(t))$ est le vecteur de covariables dépendant du temps, pour le $i^{ème}$ contrat, alors le modèle de Aalen s'écrit

$$\alpha(t; \mathbf{Z}_i(t)) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t)Z_{ik}(t) = \beta_0(t) + \beta^T(t)Z_i(t).$$

C'est un modèle additif non paramétrique, puisqu'aucune forme particulière des fonctions de régression $\beta(t)$ n'est précisée, *a priori*. L'objectif est d'estimer et de tester les coefficients $\beta_k(t)$ avec $0 \leq k \leq p$.
Une représentation graphique de ces quantités permettra d'étudier les variations de l'influence des covariables au cours du temps, sur la durée de vie des contrats.

Le modèle de Aalen s'écrit

$$E(d\underline{N}(t)/F_{t^-}) = Y(t)\beta(t)dt;$$

avec *Y*(t) matrice d'intensité (*n*,*p*+1) où chaque ligne est définie par $Y_i(t) = (1, Z_{i1}(t), ..., Z_{ip}(t)),\ 1 \leq i \leq n$ et $\beta(t) = (\beta_k(t))_{0\leq k\leq p}$.

Le théorème de Doob-Meyer, pour un processus de dénombrement multivarié, nous permet d'écrire

$$d\underline{N}(t) = d\underline{M}(t) + Y(t)dB(t),$$

où $B(t) = (B_k(t))_{0\leq k\leq p}$, avec $B_k(t) = \int_0^t \beta_k(u)du$ et $d\underline{M}(t)$ martingale vectorielle centrée.

3.2 Estimation du paramètre $B(t)$

Afin de résoudre $d\underline{N}(t) = d\underline{M}(t) + Y(t)dB(t)$, on utilise un inverse généralisé $Y^+(t)$.

En posant $J(t) = 1_{\{\text{ rang } Y(t)=p+1\}}$, l'on peut prendre $\hat{B}(t) = \int_0^t J(u)Y^+(u)d\underline{N}(u)$ comme estimateur des moindres carrés de $B(t)$, ou encore

$$\hat{B}(t) = \sum_{j:t_{(j)}\leq t} J(t_{(j)})(Y^T(t_{(j)})Y(t_{(j)}))^{-1}Y^T(t_{(j)})\Delta\underline{N}(t_{(j)}),$$

avec $\Delta\underline{N}(t_{(j)})$ vecteur dont les composantes $\Delta\underline{N_i}(t_{(j)})$ sont les indicatrices de la défaillance de l'individu i à l'instant $t_{(j)}$.

En considérant la matrice $X(t)$ à *n* lignes et *p+1* colonnes dont la $i^{ème}$ ligne est de la forme $X_i(t) = (1, Z_{i1}(t), ..., Z_{ip}(t))$, si le $i^{ème}$ individu est à risque en *t*, sinon cette ligne est formée de zéros, nous pouvons alors écrire

$$\hat{B}(t) = \sum_{j:t_{(j)}\leq t} X^T(t_{(j)})X(t_{(j)}))^{-1}X^T(t_{(j)})\boldsymbol{I}(t_{(j)},$$

avec $\boldsymbol{I}(t_j)$ vecteur $(n, 1)$ dont le $i^{ème}$ élément vaut 1 si l'individu défaillit en $t_{(j)}$; 0 sinon.

Remarque : Ayant obtenu les $\hat{B}(t)$, on déduit les $\hat{\beta}_k(t)$, avec $1 \leq k \leq p$, en considérant simplement la pente des $\hat{B}_k(t)$ ou en utilisant des méthodes du type lissage à noyau.

3.3 Tests de nullité des coefficients de régression

Soit l'hypothèse $H_0 : \beta_k(t) = 0$. Lorsque cette hypothèse est vérifiée, cela implique la nullité des coefficients de régression sur une période $[0, \tau]$; ce qui valide l'hypothèse $H_0 : B_k(t) = 0$. Martinussen et Scheike (2005) proposent d'étudier la statistique de test $T_s = \sup_{t\in[0,\tau]} \left|\hat{B}_k(t)\right|$.

Les quantiles de la statistique T_s sont difficiles à obtenir ; aussi, en tenant compte du fait que $U^{(n)} = n^{1/2}(\hat{B} - B)$ converge en probabilité vers une martingale gaussienne U, de fonction variance $\Phi(t)$ (expression dépendant de $Y_i(t)$, $\mathbf{Z}_i(t)$ et $\beta(t)$), Martinussen et Scheike indiquent que $n^{1/2}T_s$ a une distribution limite équivalente à $\sup_{t\in[0,\tau]} |\mathrm{U}_k(t)|$ où U_k est la $\mathrm{k}^{ième}$ composante de U.

Il est alors possible d'obtenir une distribution asymptotique des quantiles, en utilisant la distribution asymptotique U avec un estimateur $\hat{\Phi}(t)$ de la variance $\Phi(t)$. Cette distribution devant être simulée, en considérant alors $\hat{\Phi}(t)$.

4 Application

4.1 Les données

Les données étudiées proviennent d'une compagnie, de taille signficative sur le marché français d'assurance non-vie. L'étude concerne uniquement le portefeuille *Auto* géré par une agence et pour des raisons de confidentialité, cet extrait est non représentatif du portefeuille global. Cependant, cette sélection conserve des caractéristiques suffisamment pertinentes, pour permettre une mise en oeuvre des modèles de durée de vie et une interprétation des résultats de cette agence. Après avoir éliminé quelques valeurs aberrantes, le fichier final comporte 1461 contrats *Auto*. Par ailleurs, l'agence est choisie telle que le pourcentage des résiliations, autres qu'à l'initiative des assurés, soit négligeable. Tous les types de résiliations de contrats *Auto* ont été pris en compte.
Les contrats ont été créés durant la période allant du 13 juin 1974 au 28 décembre 1995 ; la date de résiliation est située après le 1er janvier 1996. Dans cette étude, la variable d'intérêt est la durée de vie des contrats (appelée DurVie), c'est à dire que si la date de résiliation est située avant le 7 février 2006 nous considérons la différence entre la date de résiliation et la date de de création du contrat ; sinon nous avons considéré l'écart entre le 7 février 2006 et la date de création du contrat comme une censure droite fixe.

Pour chaque contrat étudié, nous disposons de différentes variables exogènes :
• AgeVehic : l'âge du véhicule ; c'est une variable quantitative définie comme l'écart entre le 1er janvier 1996 et la date de mise en circulation du véhicule . Cette variable a été codée comme suit :
- Agevehic1 correspond à Agevehic inférieur ou égal à 1,
- Agevehic2 correspond à Agevehic compris entre 1 et 4 (inclus),

- Agevehic3 correspond à Agevehic compris entre 4 et 8 ,
- Agevehic4 correspond à Agevehic strictement spérieur à 8 ;

• BM : le Bonus-Malus ; cette variable a été codée :
- BM1 correspond à un BM = 0.5 (50% de bonus),
- BM2 correspond à un BM compris entre 0.5 et 0.7(entre 30% et 50% de bonus),
- BM3 correspond à un BM strictement supérieur à 0.7 (moins de 30%de bonus, ou bien malus) ;

• Formule : la formule d'assurance ; cette variable a été codée de la façon suivante :
- Formule1 correspond à Tierce intégrale (formule tous risques),
- Formule2 correspond à Tierce Maxi ou Tierce Collision (formule RC+dommages),
- Formule3 correspond à Tierce Simple(formule RC seule).

4.2 Statistiques exploratoires

Les tableaux suivants présentent la répartition des contrats en résiliés et censurés, leur durée de vie ; les variables exogènes sont fixées au 1er janvier 1996.

Contrats	Résiliés	Censurés
1461	927	537

TAB. 1 – *Répartition des contrats étudiés.*

	Contrats	Résiliés	Censurés
DurVie	10.24	7.59	14.79

TAB. 2 – *Durée de vie moyenne (en années).*

Remarquons que la durée de vie moyenne, de l'ensemble des contrats, est de 10.24 années. Les tableaux suivants présentent la durée de vie moyenne, selon les segments étudiés.

	Contrats	Résiliés	Censurés
BM1	12.63	9.73	15.94
BM2	9.77	7.26	14.11
BM3	7.89	6.09	13.18

TAB. 3 – *Durée de vie moyenne, selon le bonus-malus.*

4.3 Estimation du modèle de Cox

On introduit cette section, par les résultats de l'estimation des paramètres du modèle de Cox et les tests associés.

	Contrats	Résiliés	Censurés
Formule1	9.19	6.70	12.80
Formule2	10.58	7.96	14.92
Formule3	10.52	7.71	16.22

TAB. 4 – *Durée de vie moyenne, selon la formule.*

	Contrats	Résiliés	Censurés
Agevehic1	6.20	4.77	10.62
Agevehic2	8.20	5.53	11.78
Agevehic3	8.84	6.43	13.17
Agevehic4	12.16	9.25	17.34

TAB. 5 – *Durée de vie moyenne, selon la formule.*

	coef	exp(coef)	se(coef)	z	p
BM	0.419	1.520	0.0400	10.47	0.0e+000
Formule	0.181	1.199	0.0577	3.14	1.7e-0.003
Agevehic	-0.326	0.722	0.0522	-6.25	4.2e-0102

Likelihood ratio statistic= 174 on 3 df, p-value $\simeq 0$;
Wald statistic= 174 on 3 df, p-value $\simeq 0$;
Score statistic= 179 on 3 df, p-value $\simeq 0$.

TAB. 6 – *Test de l'effet des variables.*

Le modèle appris sur les données, est

$$\alpha(t;\mathbf{Z}) = \alpha_0(t)\exp(0.419\ \text{BM}\ + 0.181\ \text{Formule} - 0.325\ \text{Agevehic}).$$

L'ensemble des tests (Likelihood, Wald et Score) conduisent à la non nullité des coefficients du modèle (la p-value est inférieure à un seuil de 5% pour chacun des 3 tests).
Au vu du tableau, chacun des coefficients du modèle est significatif. Nous pouvons interpréter les valeurs des coefficients, comme suit : plus le Bonus ou la couverture diminue, plus le risque de résiliation augmente ; de même, plus l'âge du véhicule est grand, moins le risque de résiliation est important.
En d'autres termes, un faible Bonus ou une couverture faible sont des facteurs de risque, de résiliation du contrat, alors que l'augmentation de l'âge du véhicule n'en est pas.
Sous l'hypothèse de hasard proportionnel et en étudiant les *risk ratios*, on constate que le passage, d'une classe de Bonus à la suivante, augmente en moyenne le risque de résiliation de 52% (les autres variables étant maintenues fixes) puisque $\exp 0.419 = 1.52$; de même, le passage d'une classe de Formule à la suivante augmente en moyenne le risque de résiliation de 19,9% ($\exp 0.181 = 1.199$) ; et le passage d'une classe d'age du véhicule (valeur de la variable Agevehic) à la suivante diminue en moyenne le risque de 28% environ puisque $\exp(-0.326) = 0.722$ (les autres variables étant bien maintenues fixes).

4.4 Tests de l'hypothèse $(\mathcal{PH})$

4.4.1 Test graphique

Partant de la relation, valable pour les covariables à niveaux,
$\log(-\log S(t;\mathbf{Z}_1)) - \log(-\log S(t;\mathbf{Z}_2)) = \beta^T(\mathbf{Z}_1 - \mathbf{Z}_2)$,

On représente les courbes log-log, des fonctions de survie, en fonction du temps. Si l'hypothèse $(\mathcal{PH})$ est vérifiée, les courbes estimées doivent être approximativement parallèles.

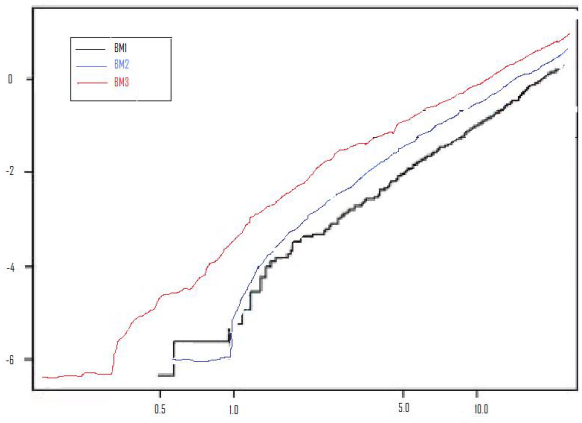

FIG. 1 – *Représentation, selon les strates de* `BM`*, de* $\log(-\log(S(t)))$*, en fonction du temps (en années).*

On constate ici que l'écart n'est pas constant entre ces courbes, que ce soit pour la covariable `BM` (figure1), ou `Agevehic` (figure2) ou bien encore `Formule`(figure3).L'hypothèse $(\mathcal{PH})$ n'est donc pas respectée sur ces différents graphiques.

4.4.2 Résidus de Schoenfeld normalisés

Les courbes lissées des paramètres $\beta(t)$ pour les covariables BM, Formule et Agevehic représentées en fonction du temps (cf figures 5,6 et7 ne correspondent pas à des droites horizontales,l'hypothèse $(\mathcal{PH})$ n'est pas valide.
Ceci est confirmé par les tests statistiques, effectués sur les pentes des droites ajustées aux courbes pour chaque covariable : chacune des pentes est significative (la p-value est inférieure à un seuil de 5% pour chacun des tests. Pour cette étude, nous avons considéré $\beta(t) = \beta + \theta t$ et nous avons testé l'hypothèse $H_0 \ : \ \theta = 0$.

4.5 Modèles alternatifs au modèle de Cox

L'hypothèse de hasard proportionnel $(\mathcal{PH})$ n'étant pas satisfaite, nous étudions comme modèles alternatifs, la partition de modèles (ou modèle de Cox stratifié) et le modèle de Aalen.

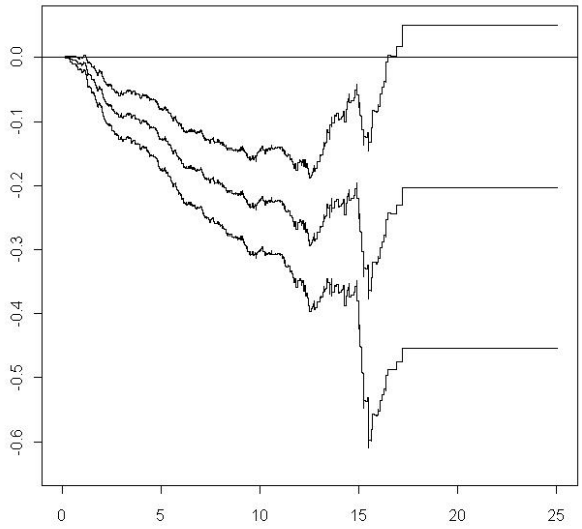

FIG. 2 – *Représentation, selon les strates de* `AgeVehic`*, de* $\log(-\log(S(t)))$*, en fonction du temps (en années).*

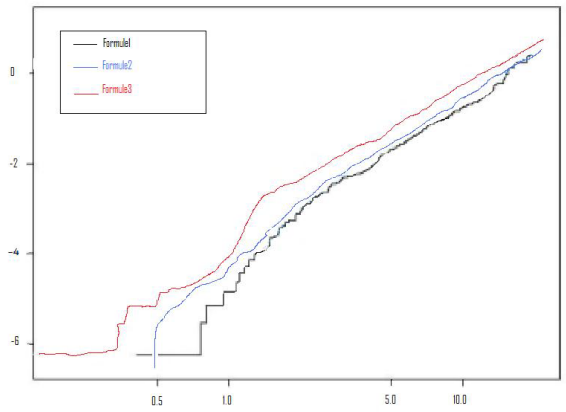

FIG. 3 – *Représentation, selon les strates de* `Formule`*, de* $\log(-\log(S(t)))$*, en fonction du temps (en années).*

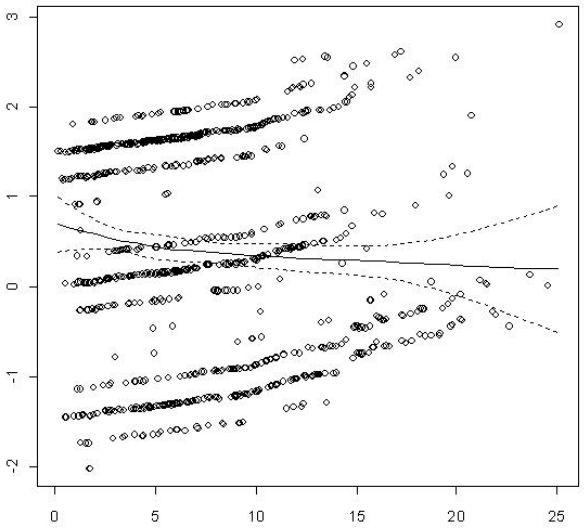

FIG. 4 – *Résidus de Schoenfeld (`BM`)*

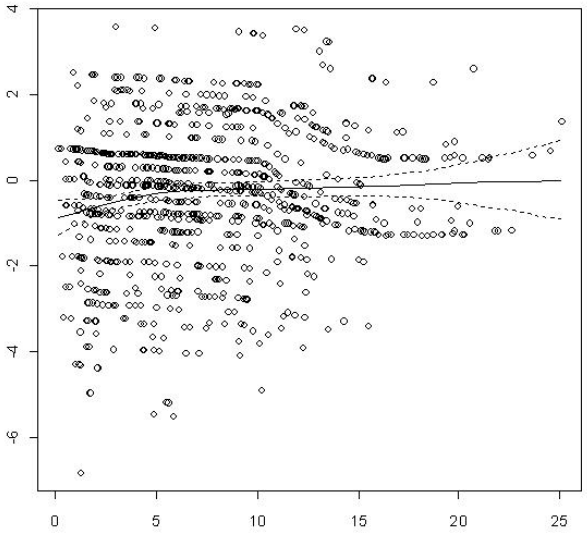

FIG. 5 – *Résidus de Schoenfeld (`Agevehic`)*

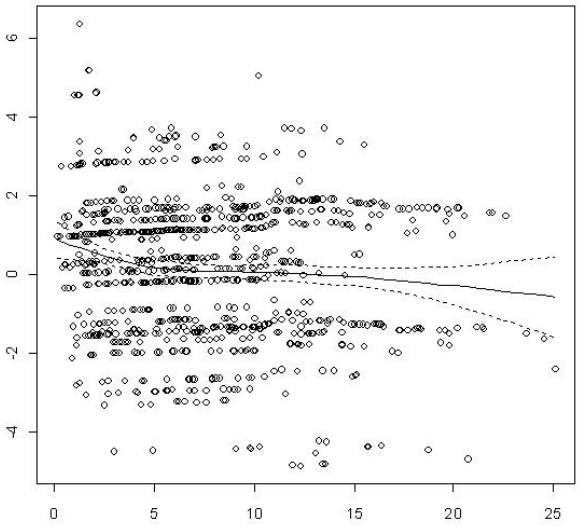

FIG. 6 – *Résidus de Schoenfeld (`Formule`)*

	rho	chisq	p
Agevehic	0.0987	9.19	2.43e-03
Formule	-0.1174	13.46	2.43e-04
BM	-0.0849	6.33	1.19e-02
Global	NA	24.10	2.38e-05

TAB. 7 – *Tests de la dépendance au temps, des coefficients β, associés aux covariables. Ici, tous les coefficients dépendent du temps.*

4.5.1 Modèles de Cox stratifiés

On étudie la survie, covariable par covariable *i.e.,* on met en oeuvre le modèle de Cox, avec stratification selon les différentes covariables.
Constatons, sur la base des figures précédentes, que un âge fixé, plus le Bonus est important ou l'âge du véhicule est grand ou encore la couverture est forte, moins la résiliation est probable.

4.5.2 Modèle de Aalen

Pour le $i^{ème}$ assuré, le modèle s'écrit

$$\alpha(t;\mathbf{Z}_i(t)) = \beta_0(t) + \sum_{k=1}^{3} \beta_k(t) Z_{ik}(t).$$

Dans le tableau suivant, on teste la nullité des coefficients, c'est à dire $H_0 : \beta_k(t) = 0$.
On constate que les coefficients sont tous significatifs (la *p-value* est inférieure à un seuil de 5% pour chacun de ces tests). L'étude graphique des fonctions de régression cumulées

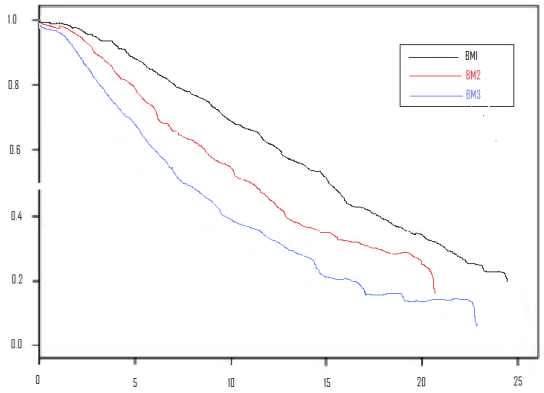

FIG. 7 – *Représentation de la survie, selon les strates de* `BM`.

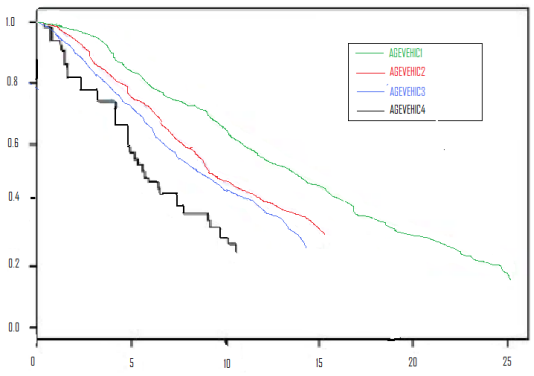

FIG. 8 – *Représentation de la survie, selon les strates de* `Agevehic`

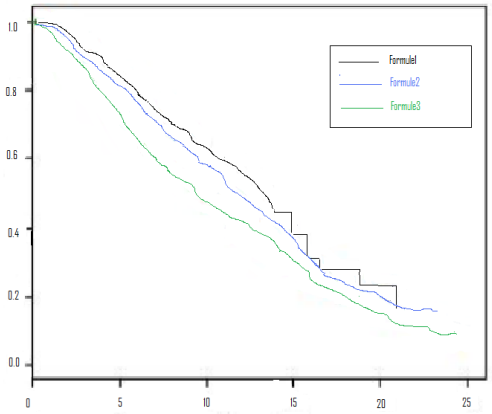

FIG. 9 – *Représentation de la survie, selon les strates de* `Formule`.

	Supremum-test	p
Intercept	4.75	0.00
Agevehic	6.17	0.00
Formule	4.79	0.00
BM	9.30	0.00

TAB. 8 – *Test de nullité des coefficients $\beta(t)$. Ici, tous les coefficients, sont significativement différents de zéro.*

$B_k(t) = \int_0^t \beta_k(u)du$, permet de voir comment les paramètres évoluent en fonction du temps.
$\widehat{B}_0(t)$, estimation de la fonction cumulée de base, croît linéairement pendant les 13 premières années. Cela suggère que le paramètre estimé $\widehat{\beta_0}$ est approximativement constant. $\widehat{B}_1(t)$ correspond à la covariable BM, il indique que plus le malus croît, plus la probabilité de résiliation est forte et ceci sur toute la période de 15 ans.
$\widehat{B}_2(t)$ correspond à la covariable Formule, ce coefficient croît linéairement jusqu'à 6 ans environ puis stagne ensuite. Ceci signifie que moins l'individu est "couvert" (formule "faible") plus sa probabilité de résiliation est forte au cours des 6 premières années, après 6 ans la probabilité de résiliation est quasiment constante.
$\widehat{B}_3(t)$ correspond à la variable `Agevehic`. Sa représentation graphique indique que plus le véhicule est ancien, plus la probabilité de résiliation décroît et ceci pendant la période $[0, 12 \text{ ans}]$, entre 12 et 15 ans la probabilité de résiliation se met à croître ; ceci pourrait s'expliquer par le fait que le véhicule devenant trop ancien, l'assuré décide de s'en séparer.

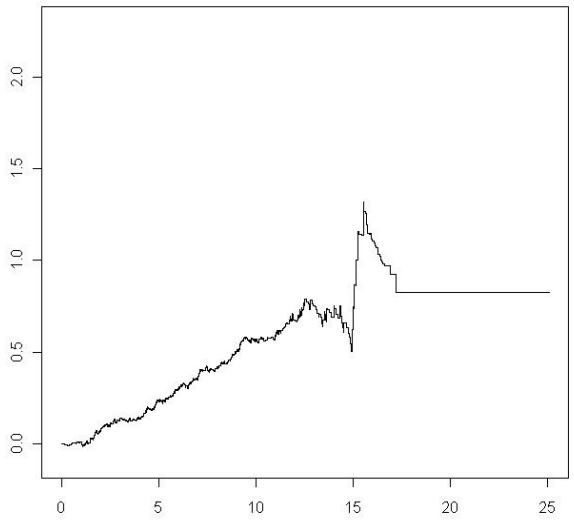

FIG. 10 – *Représentation de l'intercept cumulé, en fonction du temps.*

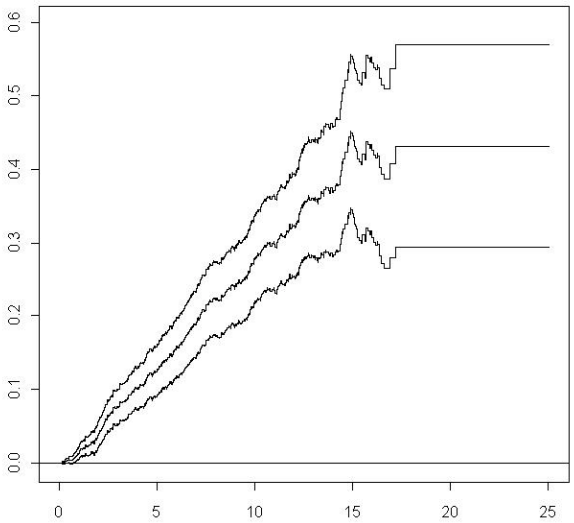

FIG. 11 – *Représentation du coefficient cumulé, associé au strates de* `BM`*, en fonction du temps.*

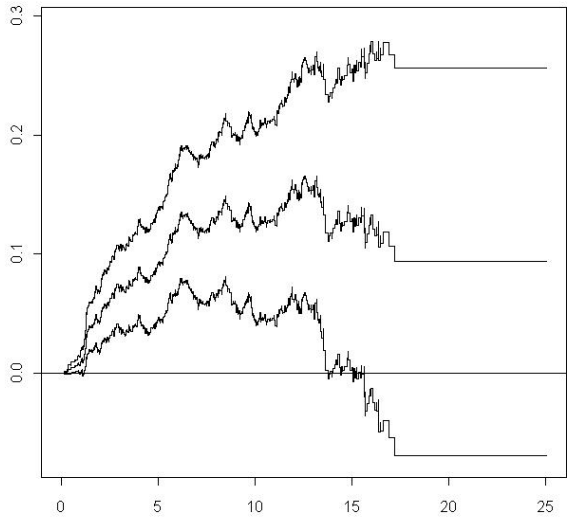

FIG. 12 – *Représentation du coefficient cumulé, associé au strates de* `Formule`*, en fonction du temps.*

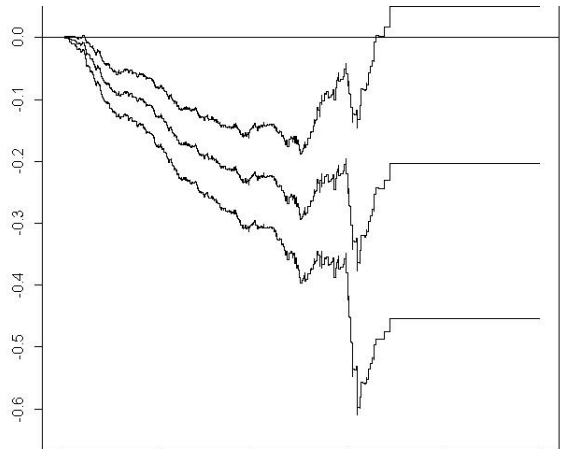

FIG. 13 – *Représentation du coefficient cumulé, associé au strates de* `Agevehic`*, en fonction du temps.*

5 Conclusion

Cette étude sur le phénomène de résiliation de contrats, a permis d'illustrer quelques alternatives au modèle de Cox, lorsque l'hypothèse de "hasard proportionnel" n'est pas vérifiée. Le modèle additif de Aalen a pu ainsi être introduit. Il permet d'étudier l'évolution des coefficients associés aux variables exogènes du modèle et ainsi, de mieux expliquer la résiliation (ou la durée de vie) des contrats.

Références

Andersen, P., O. Borgan, R. Gill, et N. Keiding (1993). *Statistical models based on counting processes*. Springer.

Cox, D. (1972). Regression model and life tables. *J.R.Stat.Soc., B 34*, 187–220.

Cox, D. et D. Oakes (1984). *Analysis of survival data*. Chapman and Hall.

Grambsch, P. et T. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*, 515–526.

Kalbfleisch, J. et R. Prentice (1980). *The statistical analysis of failure time data*. Wiley.

Klein, J. et M. Moeschberger (2003). *Survival analysis : Techniques for Censored and Truncated Data*. Springer.

Marion, J., J. Loizeau, et A. Oulidi (2007). Méthodes d'estimation de durées de vie de contrats d'assurances automobiles. *RNTI A1 : Data Mining et apprentissage statistique : applications en assurance, banque et marketing*, 157–170.

Martinussen, T. et T. Scheike (2005). *Dynamic regression models for survival data*. Springer-Verlag.

Planchet, F. et P. Thérond (2006). *Modèles de durée*. Economica.

Therneau, T. et P. Grambsch (2001). *Modeling survival data*. Springer.

Winnett, A. et P. Sasieni (2001). A note on scaled schoenfeld residuals for the proportional hazards model. *Biometrika 88*(2), 565–571.

Summary

This work discusses the methodology for the study of car insurance policies. The approach is the use of survival models, depending on feature variables or covariates. The data consist of a large database corresponding to the portfolio of contracts cars, from a large European insurance company. We test, using different approaches, the assumption of proportionality of hazard functions on which the Cox model is based. It appears from these data, the inadequacy of such a model. The alternative choice is a model with parameters as depending on time, such the Aalen model.