Apport des méthodes de durée de vie au domaine de l'assurance. Application aux contrats d'assurances automobiles

Kamal Boukhetala, Jean-Marie Marion, Abder Oulidi

Apport des méthodes de durée de vie au domaine de l'assurance. Application aux contrats d'assurances automobiles

K. Boukhetala *, J.M. Marion**, A. Oulidi**

* *USTHB - Département de Probabilités et Statistiques - Alger*
** *Institut de Mathématiques Appliquées – CREAM UCO*
44, rue Rabelais, BP 808, 49008 Angers Cedex 01

Résumé
On présentera dans cette communication les différents modèles en analyse de durée de vie utilisés dans le domaine de l'actuariat de l'assurance: modèles paramétriques, non paramétriques et semi-paramétriques. On étudiera plus particulièrement le phénomène de résiliation de contrats d'assurances automobiles d'un portefeuille d'une compagnie de taille significative sur le marché français d'assurance non-vie. En particulier nous considérerons le modèle de Cox avec covariables dépendant du temps que nous comparerons à d'autres modèles.

Mots clés *Modèle de Cox, lois de survie, données censurées, Kaplan-Meier, durées de vie de contrats d'assurances Auto.*

Abstract
In this communication we are interested in survival models and their applications on actuarial problems. We particularly study the lifespan of car's insurance contracts and the phenomenon of cancellation. Thus, we compare the lifespan of car's insurance contracts estimated by traditional survival models (nonparametric, parametric and semi-parametric). We exploit also the results obtained with the Cox model where covariates are time dependent.

Keywords *Cox model, survival distributions, , censored data, Kaplan-Meier, lifespan of car's insurance contracts.*

1- Introduction

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) [2,3,4,5]. 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 [6]. Dans ce travail, nous nous intéressons à l'estimation des durées de vie de contrats d'assurance automobile afin de mieux gérer le phénomène de résiliation de contrats.

En France, l'assurance automobile est un marché mature avec un faible taux de croissance. De plus, s'agissant d'un secteur convoité, de nouveaux intervenants (les banques-assureurs, les acteurs de la grande distribution …) viennent rejoindre les acteurs traditionnels. Confrontés à

une forte concurrence exacerbée par la quasi-stabilité du parc automobile assurable, et face aux mutations importantes de leur référentiel comptable et réglementaire, les assureurs sont désormais incités, plus qu'avant, à développer des modèles optimaux de surveillance et de gestion de leur portefeuille afin, entre autres, de fidéliser les clients les plus rentables et éventuellement de résilier certains contrats.
Les analyses d'ordre qualitatif s'avèrent rapidement très insuffisantes, et les assureurs ont recours, de plus en plus, à des techniques statistiques plus élaborées. Des techniques de classification, des modèles de prévisions permettent de réaliser la tarification et d'élaborer des modèles prédictifs de résiliation de certains contrats d'assurance automobile.

Dans ce travail on s'intéressera tout particulièrement à l'étude du phénomène de résiliation et à l'élaboration d'un modèle prédictif de durée de vie de contrats Auto. Une application pratique sera réalisée sur un extrait du portefeuille d'une compagnie de taille significative sur le marché français d'assurance non-vie. On entend par durée de vie de contrat Auto, la durée séparant la date de résiliation de la date de création de contrat. La date de création de contrat est connue, par contre la date de résiliation 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.

En effet, les modèles de survie sont adaptés dès que les phénomènes d'intérêt se modélisent comme des variables aléatoires positives avec éventuellement des données manquantes. Ils sont utilisés dès qu'on cherche à modéliser et à estimer les lois décrivant le temps qui s'écoule entre deux événements à partir d'observations de durées et éventuellement de variables explicatives dites exogènes (ou covariables).

Désignons par *T* une variable d'intérêt , c'est à dire une variable aléatoire positive décrivant le temps qui s'écoule entre deux évènements : par exemple la durée de vie d'un contrat qui peut être définie comme la différence entre la date de résiliation et la date de création du contrat.

Nous supposerons que la distribution de T possède en tout point une densité de probabilité *f*, sa fonction de répartition sera notée *F*.

Les fonctions utilisées habituellement en analyse des données de survie sont :

- la fonction de survie *S* définie par : $S(t)=P(T>t)$
- la fonction de hasard *h* définie par : $h(t)=\frac{f(t)}{S(t)}$, qui est une caractéristique locale

s'interprétant au point *t* comme la probabilité instantanée de sortie de l'état sachant que le sujet est encore dans cet état à l'instant *t*, soit : $h(t)=\lim_{\Delta t\to 0}\frac{1}{\Delta t}P\left(T\in\left]t,t+\Delta t\right]/T>t\right)$

- la fonction de hasard intégrée *H* définie par : $H(t)=\int_0^t h(x)dx$

notons que : $S(t)=\exp(-H(t))$ puisque $S(0)=1$.

Ces fonctions caractérisent entièrement la distribution de *T*. Nous nous intéressons dans la suite au problème de l'estimation de la fonction de survie.

Ce n'est pas toujours *T*, la variable d'intérêt qui est observée, mais une autre variable *C*, appelée *censure* qui donne une information sur *T*.

Par exemple, si nous ne connaissons pas la date de résiliation du contrat : au lieu de *T*, nous observons *(Y,D)* avec $Y = \inf(T, C)$ et D indicatrice de l'événement $\{T \leq C\}$, si *D=0* nous dirons dans ce cas que *T* est *censuré à droite*.

2- Les modèles

2.1- Les modèles non paramétriques

Les modèles non paramétriques permettent d'estimer l'une des différentes fonctions caractérisant la distribution de la variable *T* sans faire aucune hypothèse a priori sur celle-ci.

Considérons un échantillon $(T_1, ..., T_n)$ de la variable d'intérêt *T*, nous observons en réalité $Y_i = \inf(T_i, C_i)$ avec $D_i = 1_{\{T_i \leq C_i\}}$ indicatrice de l'événement $\{T_i \leq C_i\}$ où les C_i $(1 \leq i \leq n)$ désignent les censures droites.
Soit $Y_{(1)}, ..., Y_{(n)}$ la statistique d'ordre associée à $Y_1, ..., Y_n$ et $D_1^{'}, ..., D_n^{'}$ les indicatrices ordonnées correspondantes, nous définissons *R(t)* le nombre des « individus à risque » à l'instant *t*, c'est à dire par exemple les contrats qui sont encore présents à t^- (ni résiliés, ni censurés) et nous notons $M(Y_{(i)})$ le nombre de résiliation à $Y_{(i)}$.
L'estimateur le plus utilisé est celui de Kaplan-Meier qui se définit de façon générale par :

$$\hat{S}(t) = \prod_{\{i; Y_{(i)} < t\}} \left(1 - \frac{M(Y_{(i)})}{R(Y_{(i)})}\right)$$

L'estimateur de Kaplan-Meier est un estimateur qui permet de tenir compte des censures, il est consistant et asymptotiquement gaussien mais présente l'inconvénient d'être biaisé et par nature discontinu.

2.2- Les modèles paramétriques :

Le choix d'un modèle paramétrique suppose que la loi de probabilité de la durée de vie *T* appartient à une classe de distributions de type connu, fonction de paramètres dont l'objectif sera de les estimer à partir d'un ensemble d'observations.

Considérons un échantillon $t_1, .., t_n$ de durées issu d'une distribution connue de densité $f(x, \theta)$ où θ est un paramètre inconnu qui peut être vectoriel.
En fait, nous observons $y_1, ..., y_n$ où certaines valeurs sont des censures droites, d'autres des censures gauches, enfin certaines valeurs correspondent aux t_i.

Les modèles paramétriques permettent l'introduction de variables exogènes ; par exemple on peut penser que la durée de vie d'un contrat d'assurance Auto est liée au Bonus-Malus, au type de contrat, à l'âge du véhicule …

Parmi les modèles les plus répandus, nous allons considérer le modèle de régression log linéaire défini par

$$\ln y_i = \tilde{x}_i \beta + \eta \varepsilon_i \qquad \text{pour } 1 \leq i \leq n$$

avec :

$\tilde{x}_i$ transposé du vecteur des variables exogènes correspondant à l'individu *i*

β vecteur des paramètres

η paramètre d'échelle, par défaut sa valeur est prise égale à 1

ε_i variable aléatoire dont la distribution de probabilité est connue (exponentielle, Weibull, log logistique,…)

Les paramètres η, β et ceux de la loi des ε_i sont estimés par la méthode du maximum de vraisemblance.

Connaissant la loi des ε_i, on en déduit la fonction de survie de *T*.

2.3- Les modèles semi-paramétriques :

Les modèles semi-paramétriques cherchent à estimer la fonction de survie en tenant compte de l'influence des facteurs exogènes et sans faire aucune hypothèse a priori sur la forme de la distribution de base.
Parmi ces modèles, le plus utilisé est *le modèle à hasard proportionnel de Cox* [1].
La fonction de hasard, s'écrit dans ce cas :

$$h(t/x) = h_0(t) \exp(\tilde{x}\beta)$$

où :

- h_0 est la fonction de hasard de base (non spécifiée) - elle peut s'estimer non paramétriquement,
- $\tilde{x}$ est le transposé du vecteur des variables exogènes,
- β est le vecteur des paramètres à estimer.

Si l'on note S_0 la fonction de survie de base associée à h_0, on a la relation suivante :

$S(t/x) = \left[S_0(t)\right]^{\exp(\tilde{x}\beta)}$, ce qui permet d'obtenir une estimation de *S* connaissant l'estimation du vecteur β.

Pour estimer les composantes du vecteur β, à partir d'un échantillon ordonné $\left(y_{(1)},...,y_{(n)}\right)$, on calcule la fonction de vraisemblance partielle de Cox qui n'est autre que (s'il n'y a pas de données censurées) :

$$L\left(y_{(1)},...,y_{(n)};\beta\right) = \prod_{i=1}^{n} \frac{\exp(\tilde{x}_i\beta)}{\sum_{k \in R\left(y_{(i)}\right)} \exp(\tilde{x}_k\beta)}$$

Cette fonction de vraisemblance partielle reste identique dans le cas de données censurées à droite.

Notons que dans ce cas, une estimation de la fonction de survie peut être donnée par la relation suivante :

$$\hat{S}(t/x) = \prod_{\{j;y_{(j)}<t\}} \hat{\upsilon}_j^{\exp(\tilde{x}\hat{\beta})}$$

où les $\hat{\upsilon}_j$ sont solutions des équations de vraisemblances :

$$\sum_{l\in D_{3_j}} \frac{\exp(\tilde{x}_l\hat{\beta})}{1-\upsilon_j^{\exp(\tilde{x}_l\hat{\beta})}} = \sum_{l\in R(y_{(j)})} \exp(\tilde{x}_l\hat{\beta}) \quad \text{pour } 1\le j\le z$$

dans lesquelles z est le nombre de durées de vie distinctes et D_{3_j} correspond aux durées de vie effectivement observées dans l'échantillon.

Le modèle de Cox peut être généralisé en considérant non pas que les variables exogènes restent fixes pendant toute la durée d'observation mais qu'elles évoluent au cours du temps (par exemple la variable Bonus-Malus…)
Ceci se traduit par une information supplémentaire pour évaluer les durées de vie.
Connaissant l'évolution des $\tilde{x}_i$ en fonction du temps, on évalue alors les termes qui entrent dans la vraisemblance partielle de Cox de la façon suivante :

$$L(y_{(1)},...,y_{(n)};\beta) = \prod_{i=1}^{n} \frac{\exp(\tilde{x}_i(y_{(i)})\beta)}{\sum_{k\in R(y_{(i)})} \exp(\tilde{x}_k(y_{(k)})\beta)}$$

On obtient une nouvelle estimation du vecteur β et par suite une estimation plus précise de la fonction de survie.

3- Application à des données de contrats d'assurances automobiles

Les données analysées sont issues d'un portefeuille provenant d'une compagnie de taille significative sur le marché français d'assurance non-vie. Après avoir éliminé quelques valeurs aberrantes[1], le fichier final que nous étudions comporte 1557 contrats Autos.

Dans ce papier, l'objectif est de présenter des méthodes pour estimer la durée de vie des contrats d'assurance Auto. Tous les types de résiliations de contrats Autos ont donc été pris en compte, qu'elles soient à l'initiative du client ou de la compagnie (suite à disparition du risque, vente ou fin de vie du véhicule suite panne ou accident, pour non-paiement, ou encore d'un commun accord entre assureur et assuré …). Bien entendu, ces différentes résiliations ne sont pas homogènes et peuvent avoir des résultats sensiblement éloignés en terme de durée de vie du contrat pour un portefeuille dans lequel existe toutes les formes de résiliation.

Ces contrats ont été crées entre le 13 juin 1974 et le 28 décembre 1995, leur date de résiliation est située après le 1er janvier 1996. Dans l'étude qui a été faite, la variable d'intérêt est la durée de vie des contrats, c'est à dire que si la date de résiliation est située avant le 31

[1] Les valeurs aberrantes concernent uniquement quelques dizaines de contrats pour lesquels la variable date de mise en circulation du véhicule n'est pas exploitable.

décembre 2000 nous avons considéré la différence entre la date de résiliation et la date de création sinon nous avons considéré l'écart entre le 31 décembre 2000 et la date de création du contrat (nous considérons donc une censure droite fixe).

Les variables exogènes considérées sont : l'âge du véhicule, le Bonus-Malus, la formule d'assurance.

4- Résultats et Conclusion

En l'absence d'information a priori sur la forme de la fonction de survie, nous l'avons estimée d'abord par la méthode non-paramétrique de Kaplan-Meier. Puis, pour introduire des variables exogènes dans le modèle, nous avons étudié des méthodes paramétriques (modèles de régression log-linéaire) ou la loi log-logistique est apparue la mieux adaptée à nos données pour la modélisation de la durée de vie des contrats Auto. Enfin un modèle semi-paramétrique de Cox a été utilisé en stratifiant sur différentes variables exogènes lorsque les hypothèses liées au modèle le permettaient. L'impact temporel de certaines variables (Bonus-Malus) nous a permis en utilisant le modèle de Cox, de préciser l'évolution de la durée de vie de ces contrats

En conclusion, cette étude sur le phénomène de résiliation de contrats 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.

Par suite, l'estimation des durées de vie de contrats Auto peut être utilisée pour évaluer l'amortissement des coûts d'acquisition d'un contrat, mais aussi pour calculer la rentabilité d'un contrat tenant compte de sa durée de vie prévue (on obtient ainsi un des éléments de la valeur du client), et enfin tous ces éléments peuvent être analysés de manière segmentée en fonction de critères pour affiner le ciblage marketing, les tarifs…

5- Références

[1] COX D.R. et OAKES D. (1984), *Analysis of survival data*, London, Edition Chapman and Hall.

[2] DROESBEKE J.J, FICHET B, TASSI P.,éditeurs (1989), *Analyse statistique des durées de vie: Modélisation et données censurées*, Economica.

[3] KALBFLEISH J.D. et PRENTICE R.L. (1980), *The statistical analysis of failure time data*, New York: Wiley and Sons, Inc.

[4] KAPLAN E.L. et MEIER P. (1958), *Non parametric estimation from incomplete observations*, J. Amer. Statist. Assoc. 53, pp 457-481.

[5] LI, S.,(1996). *Survival analysis*, Marketing Research, 7(4), 17-23.

[6] PLANCHET F. et THEROND P. modèles de durée. Applications actuarielles. Economica (2006)