Modèle de Cox avec des données hétérogènes

Eliz Peyraud [1,2], Julien Jacques [1], Guillaume Metzler [1] et Alexandre Lopez [2]

[1] *Univ Lyon, Univ Lyon 2,*
ERIC UR 3083, Lyon, France
{eliz.peyraud,julien.jacques,guillaume.metzler}@univ-lyon2.fr.
[2] *Institut Georges Lopez (IGL)*
Lissieu, France
{epeyraud,alopez}@igl-transplantation.com

Résumé. Dans un contexte médical, l'analyse de survie se fait essentiellement à l'aide de modèles à risques proportionnels de Cox. Le présent papier cherche à construire un tel modèle afin de prédire la probabilité de survie d'un patient à un temps t après un acte chirurgical lourd. Nous mettons en évidence la nécessité de prendre en compte l'hétérogénéité des patients lors de la modélisation et expliquons comment nous pouvons prendre en compte des variables explicatives de natures variées.

Mots-clés. Modèle de Cox, Analyse de survie, Données médicales, Données hétérogènes.

Abstract. In a medical context, survival analysis is mainly done using Cox proportional hazards models. The present paper seeks to build such a model in order to predict the probability of survival of a patient at a time t after a heavy surgical procedure. We highlight the need to take into account the heterogeneity of the patients during the modeling and explain how we can take into account explanatory variables of various natures.

Keywords. Cox Proportional Hazards Model, Survival Analysis, Medical Data, Heterogeneous Data.

1 Introduction

L'analyse de survie est une méthode couramment utilisée dans le domaine médical pour, entre autre, estimer la durée de vie d'un patient suite à une opération lourde ou à la prise d'un traitement conséquent. Elle peut également se retrouver pour étudier la durée de vie d'un composant suite à une modification effectuée sur ce dernier. Dans un contexte exclusivement médical, ces modèles sont importants pour guider les médecins dans le choix des traitements à prescrire en estimant le taux de survie du patient avec la prise en compte de ce nouveau facteur.

A ce titre, les modèles de Cox [6] constituent les principaux outils de modélisation de ce phénomène via la prise en compte d'informations relatives aux patients. En revanche, l'hétérogénéité des profils de patients (représentés par les différentes covariables) peut souvent mettre à mal l'efficacité de tels modèles.

Dans cette étude, nous montrons la nécessité de prendre en compte cette hétérogénéité des données dans l'apprentissage d'un modèle de Cox. Plus précisément, nous montrons que les courbes de survie peuvent fortement varier selon la typologie du patient (*e.g.* distinction homme/femme, antécédant médicaux, origine ethnique, ...). Enfin, nous montrons également, sur un ensemble restreint de descripteurs, que leur significativité peut dépendre de cette typologie de patients.

Ces observations sont à l'origine de la perspective présentée à la fin de ce travail.

2 Analyse de Survie

On cherche à analyser la probabilité de survie du patient après un évènement ayant eu lieu à un temps t_0 (*e.g.* un acte chirurgical lourd). Cela passe par l'étude d'une variable aléatoire T qui représente la *durée de survie* du patient à partir de l'évènement qui s'est produit à t_0. On définit alors la *fonction de survie* d'un individu X, notée $S(t \mid X)$, comme la probabilité que le temps de survie du patient soit supérieur à $t > 0$:

$$S(t \mid X) = \mathbb{P}(T \geq t \mid X),$$

où X représente l'ensemble des covariables de l'individu concerné.

Deux approches historiques permettent de représenter cette fonction de survie. Une première, non-paramétrique, basée sur la courbe de Kaplan-Meier et qui ne permet pas de prendre en compte des covariables, et une deuxième, paramétrique cette fois, reposant sur les modèles de Cox.

2.1 Approche non paramétrique : Kaplan-Meier

Une première estimation non-paramétrique de la fonction de survie $S(t)$ est donnée par l'estimateur de Kaplan-Meier :

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

où t_i ($i \geq 1$) représente les instants où surviennent des évenements (une défaillance ou une censure), d_i représente le nombre d'évènements qui se sont produits jusqu'à l'instant t_i, et n_i le nombre d'individus ayant survécus jusqu'au temps t_i (y compris les individus censurés).

La censure intervient lorsque l'évènement n'a pas été observé durant le temps de l'étude. Cela peut signifier que l'évènement ne s'est pas produit, ou que l'individu est sorti de l'étude avant que celui-ci se produise. Nous faisons ici l'hypothèse que la censure est non-informative : le mécanisme de censure est indépendant de l'évènement observé. Pour prendre en compte la censure sur cet estimateur, lorsque l'on rencontre un individu censuré, l'effectif n_i au premier temps t_i après la date de censure diminue de 1.

La courbe de Kaplan-Meier a une forme caractéristique en escalier. Cette approche non-paramétrique ne tient cependant pas compte des attributs des différents observations, or ces dernières peuvent se révéler importantes dans l'estimation de la probabilité de survie d'un patient.

2.2 Modèle à risques proportionnels de Cox

Le modèle de Cox [1, 6] est un modèle semi-paramétrique servant à modéliser le phénomène d'apparition d'évènements (comme la mort d'un patient) au cours du temps, en fonction des covariables.

Ces phénomènes interviennent selon une fonction $\lambda(t, X)$, appélée *fonction de risque instantané* ou *taux de défaillance*, qui selon le modèle de Cox peut être modélisée comme suit :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta^t X)$$

où λ_0 est appelé la fonction de risque de base (autrement dit le risque instantané de l'évènement lorsque toutes les covariables sont nulles), $\beta \in \mathbb{R}^p$ est un vecteur de paramètres et $X \in \mathbb{R}^p$ un vecteur de covariables.

Dans cette expression le premier terme $\lambda_0(t)$ dépend uniquement du temps tandis que le second terme $\exp(\beta^T X)$ ne dépend que des covariables.

Une hypothèse majeure qui en découle pour pouvoir utiliser le modèle de Cox est l'hypothèse des risques proportionnels : le rapport de risque doit être constant en temps.

A partir de l'observation d'un échantillon $(X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)$, où δ_i est l'indicatrice de censure de l'individu i ($\delta_i = 0$ si l'individu est censuré, 1 sinon), l'estimateur $\hat{\beta}$ est obtenu en maximisant la log-vraisemblance partielle [5] :

$$\log\left(L(\beta, X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)\right) = \sum_{i=1}^{n} \delta_i \left[\beta^T X_i - \log\left(\sum_{j \in R(t_i)} \exp(\beta^T X_j)\right)\right]$$

où $R(t_i)$ est l'ensemble des individus susceptibles de subir une défaillance (ou censure) au temps t_i, i.e. l'ensemble des individus encore en vie au temps $t_i - \epsilon$ (ou non censurés).

Cette estimation nous permet alors de retrouver la fonction de survie associée. En effet, la fonction de survie est liée à la fonction de risque instantané par la relation suivante :

$$S(t \mid X) = \exp\left(-\int_0^t \lambda(u, X)\, du\right)$$

3 Expériences

Les données utilisées pour les expériences suivantes sont issues d'une base américaine qui retrace le suivi des patients post-opératoire (500 000 patients environ). Ces données

contiennent des informations relatives aux patients : comme la description des individus ou encore des indicateurs biologiques aux différents moments du suivi (on dénombre environ 1000 indicateurs) sur une échelle de temps pouvant varier. Les variables descriptives sélectionnées pour cette étude sont toutes de nature quantitatives.

3.1 Protocole expérimental

L'objectif de ce travail préliminaire est de démontrer la nécessité de construire des analyses de survies par sous-groupes homogènes de patients. Nous nous sommes intéressés à l'ensemble des patients sans restriction. Nous cherchons à estimer le temps de survie après opération (l'évènement observé étant donc la mort du patient). Dans notre cas, 70% de nos données étaient censurées, et les dates d'opérations étaient différentes pour chaque patient : nous avons donc calculé le temps de survie comme le temps écoulé entre la date d'opération et la date de dèces si l'évènement était observé, la date de dernier suivi du patient sinon (avec l'indicatrice de censure δ_i).

Dans un premier temps nous avons tracé les courbes de Kaplan-Meier [1]. Ensuite nous avons testé le modèle de Cox sur nos données.

Notre cohorte étant relativement grande et hétérogène, nous pouvons envisager que plusieurs populations se distinguent à l'intérieur de celle-ci. Pour confirmer cela, nous avons séparé nos observations en plusieurs groupes de population. Par exemple, nous savons que les critères d'opérations ne sont pas exactement les mêmes en fonction du genre du patient ou encore de son groupe d'âge (enfant/adulte/personne âgée).

Nous avons séparé notre population globale en plusieurs groupes successivement et de manière indépendante : séparation de la population en deux groupes selon le genre (Homme/Femme), selon la présence ou l'absence de diabète, selon l'ethnie (latine ou non), et selon le groupe d'âges (5 ans et moins, 6-17 ans, 18-49 ans, 50 ans et plus). Nous avons tracé, pour chaque groupe, les courbes de survie de Kaplan-Meier correspondantes dans le but de les comparer graphiquement.

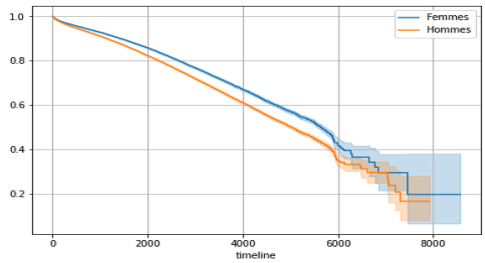
De plus, sur chacun de ces groupes nous avons également testé un modèle de Cox à 7 covariables (identiques pour chaque groupe) de nature quantitatives de manière à comparer la significativité des paramètres du modèle.

3.2 Résultats

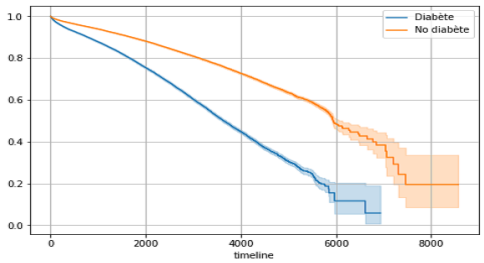
Les courbes de survie de Kaplan-Meier (probabilité de survie en fonction du temps en jours, estimée par l'approche de la section 2.1) obtenues sur nos données, et représentées en Figure 1, confirment l'existence de distinction entre les groupes étudiés en terme de probabilité de survie.

Si l'on s'intéresse en particulier à la Figure 1b on remarque que la fonction de survie décroit bien plus rapidement pour les personnes atteintes de diabète. Cette observation

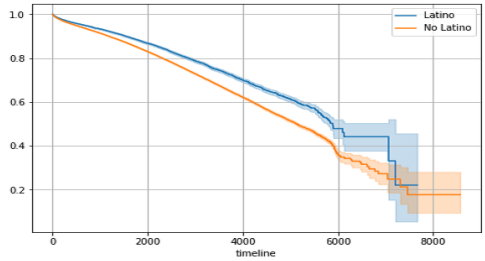
1. les courbes sont obtenues avec la librairie *lifelines* sous Python.



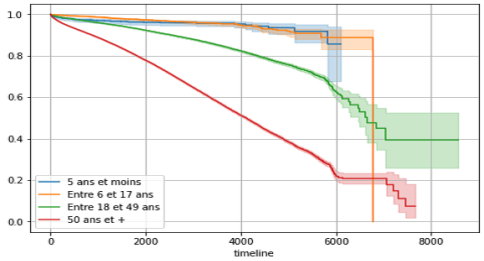
(a) Courbes de survie en fonction du genre (Homme/Femme)



(b) Courbes de survie en fonction du diabète (Présence/Absence)



(c) Courbes de survie en fonction de l'ethnie (Latine ou non)



(d) Courbes de survie en fonction de l'âge (5 ans et moins / 6-17 ans / 18-49 ans / 50 ans et plus

FIGURE 1 – Courbes de Kaplan-Meier avec intervalle de confiance à 95% sur différents sous-groupes de la population((a) Genre ; (b) Diabète ; (c) Ethnie ; (d) Âge).

rejoint les études menées en France mettant en lumière les difficultés d'obtenir les soins médicaux adéquats chez les personnes diabétiques [3]. De même, on constate également un écart important entre les courbes de survie selon les différents groupes d'âge. Ainsi, ces expériences préliminaires démontrent qu'il existe des sous-groupes de patients pour lesquels les courbes de survies sont très différentes.

	Modèle 5 ans et moins	Modèle 6-17 ans	Modèle 18-49 ans	Modèle 50 ans et plus
Âge (en mois)	0.04	0.07	< 0.005	<0.005
Creatinine donneur	0.52	0.97	0.17	0.08
Créatinine receveur	0.65	<0.005	<0.005	<0.005
Opérations antérieures	<0.005	<0.005	<0.005	<0.005
Creatinine élevée chez le donneur	0.59	0.95	0.69	0.22
Episodes de rejet aigu	0.44	0.27	<0.005	<0.005
Creatinine à la sortie de l'hôpital	<0.005	0.01	<0.005	<0.005

TABLE 1 – p-valeurs des covariables utilisées dans le modèle de Cox en fonction du groupe d'âges. *La creatinine est un déchet naturel de l'organisme. Lorsque la capacité de l'organisme à éliminer les déchets diminue, le taux de creatinine dans le sang augmente.*

En se concentrant sur la classe d'âge des patients, nous avons également réalisé quatre modèles de Cox distincts par groupes d'âges (en gardant les mêmes covariables). Ici encore les modèles pour chaque groupes se distinguent, en particulier par le nombre et le choix des variables significatives dans chaque modèle (Table 1). Par exemple, le taux de créatine du patient n'est une variable siginificative qu'à partir de l'âge 6 ans.

4 Conclusion et Perspectives

Les résultats précédents nous confirment qu'un modèle unique ne peut convenir à l'ensemble de notre jeu de données. Les modèles existants aujourd'hui, basés sur la régression de Cox classique, sont limités en ne représentant qu'une partie de la population (adultes, sans antécédants médicaux, etc). Ces critères restrictifs sont identifiés manuellement par les médecins [2].

Pour palier ce problème nous voulons détecter automatiquement des sous-groupes homogènes de la population d'un point de vue de la modélisation de la survie. Pour cela, la suite de ce travail consistera à considérer un mélange de modèles de Cox [4] :

$$\lambda(t, X) = \sum_{k=1}^{K} \pi_k \lambda_{0,k}(t) \exp(\beta_k^t X).$$

Références

[1] Breslow N. E. Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43(1) :45–57, 1975.

[2] Foucher Y. et al. A clinical scoring system highly predictive of long-term kidney graft survival. *Kidney international*, 78(12) :1288–1294, 2010.

[3] Imhoff O. et al. «le receveur limite» : existe-t-il encore des freins à l'inscription des patients sur liste d'attente de transplantation rénale ? *Néphrologie & thérapeutique*, 3 :282–288, 2007.

[4] Rosen O. and Tanner M. Mixtures of proportional hazards regression models. *Statistics in Medicine*, 18(9) :1119–1131, 1999.

[5] Cox D. R. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202, 1972.

[6] Cox D. R. Partial likelihood. *Biometrika*, 62(2) :269–276, 1975.