Modèle de Cox avec des données hétérogènes

Eliz Peyraud, Julien Jacques, Guillaume Metzler, Alexandre Lopez

Modèle de Cox avec des données hétérogènes

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

[1] Université de Lyon, Lyon 2, ERIC UR3083

{eliz.peyraud, julien.jacques, guillaume.metzler}@univ-lyon2.fr

[2] Institut Georges Lopez (IGL)

{epeyraud,alopez}@igl-transplantation.com

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.
Les **modèles de Cox** constituent les principaux outils de modélisation de ce phénomène via la prise en compte d'informations relatives aux patients.
L'hétérogénéité des profils de patients (représentés par les différentes covariables) peut cependant 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.

Approche non paramétrique

- On cherche à estimer la fonction de survie : $S(t \mid X) = \mathbb{P}(T \geq t \mid X)$
- Estimateur de Kaplan Meier :

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

$(t_i)_{i \geq 1}$ représente les instants où surviennent des évènements, 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 survécu jusqu'au temps t_i.

Approche semi-paramétrique

- Modèle de Cox [4] :

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

λ est la fonction de risque instantané, λ_0 est appelée fonction de risque de base, $\beta \in \mathbb{R}^p$ est un vecteur de paramètres et $X \in \mathbb{R}^p$ un vecteur de covariables.

- On cherche à maximiser la log vraisemblance partielle [1,5] :

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

- Lien entre fonction de risque et fonction de survie :

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

Données et expériences

Estimation du temps de survie des patients après opération

- **500 000 patients issus d'une cohorte américaine**
- informations relatives aux patients : description des individus, indicateurs biologiques aux différents moments du suivi (environ 1000 indicateurs)
- **Données censurées**

Censure : *correspond à la sortie du patient du programme de suivi (avant son potentiel décès). Nous faisons ici l'hypothèse que la censure est non-informative : le mécanisme de censure est indépendant de l'évènement observé.*

Résultats (graphiques)

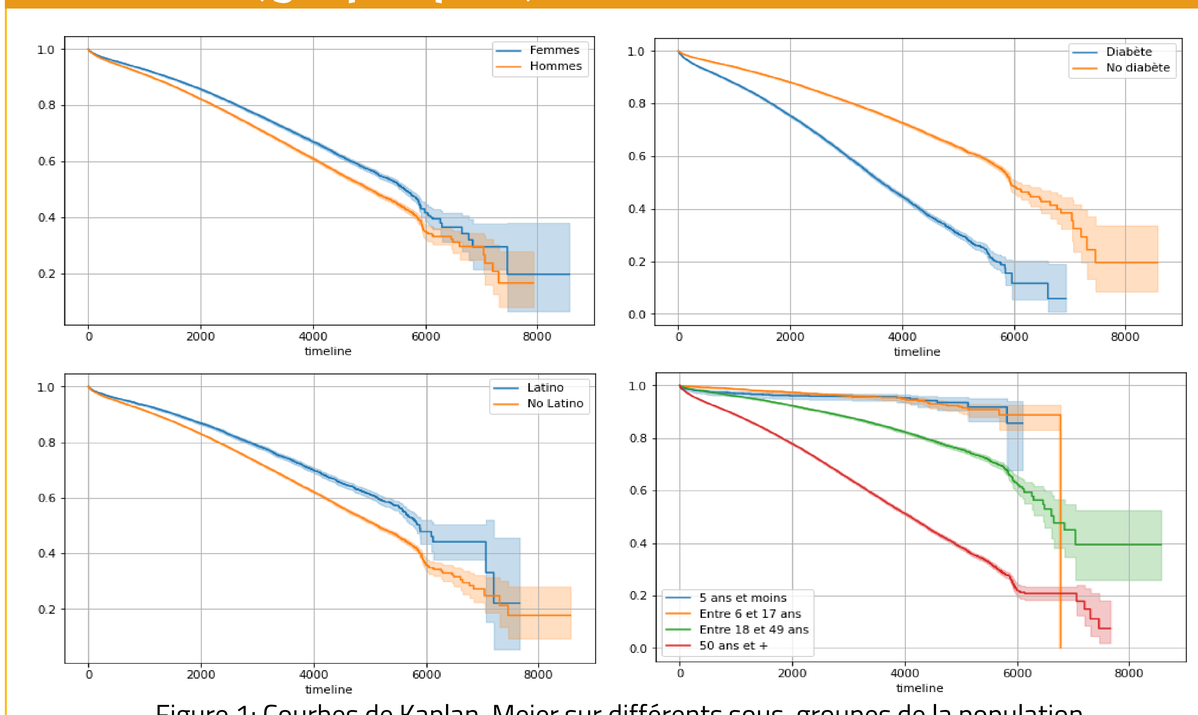

Figure 1: Courbes de Kaplan-Meier sur différents sous-groupes de la population ((a) Genre ; (b) Diabète ; (c) Ethnie ; (d) Âge). Le temps est exprimé en jours.

Résultats (graphiques)

- Nous avons calculé le temps de survie comme le temps écoulé entre la date d'opération et la date de décès si l'évènement était observé, la date de dernier suivi du patient sinon. Pour chaque observation nous avons ajouté une variable censure valant 0 si la donnée était censurée, 1 sinon.
- Nous avons tracé les courbes de Kaplan-Meier selon différents groupes étudiés : Genre (Hommes/Femmes), Diabète, Ethnie, et Âge (5 ans et moins, 6-17ans, 18-49 ans, 50 ans et plus). Les résultats sont présentés en Figure 1.

Choix du modèle

Nous avons séparé notre jeu de données en plusieurs groupes (définis en Figure 1) et nous avons testé sur chaque groupe un modèle de Cox à 7 covariables (identiques pour chaque groupe) de nature quantitatives de manière à **comparer la significativité des paramètres** du modèle.

	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.

Ainsi, par exemple, le taux de créatine du patient n'est une variable significative qu'à partir de l'âge 6 ans.

Les résultats présentés dans la Table 1 nous confirment qu'un modèle unique ne peut convenir à l'ensemble de notre jeu de données.

Conclusion et Perspectives

- Les modèles existants aujourd'hui [2] , basés sur la régression de Cox classique, sont limités en ne représentant qu'une partie de la population (adultes, sans antécédents médicaux, etc). Ces critères restrictifs sont identifiés manuellement par les médecins.
- Nous souhaitons créer un **modèle unique** prenant en compte l'hétérogénéité de la population, tout en recherchant automatiquement les différents sous-groupes à l'aide d'un **mélange de modèles de Cox** de la forme [3] :

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

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] Rosen O. and Tanner M. Mixtures of proportional hazards regression models. Statistics in Medicine, 18(9) :1119–1131, 1999.
[4] Cox D. R. Regression models and life-tables. Journal of the Royal Statistical Society : Series B (Methodological), 34(2) :187–202, 1972.
[5] Cox D. R. Partial likelihood. Biometrika, 62(2) :269–276, 1975

ERIC Entrepôts, Représentation & Ingénierie des Connaissances — Université de Lyon — Université Lumière Lyon 2 — IGL