Longitudinal data analysis : examen pratique

Les données sont issues de *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone* (Davide Chicco & Giuseppe Jurman dans BMC Medical Informatics and Decision Making, 2020) dont voici le résumé :

"Les maladies cardiovasculaires tuent environ 17 millions de personnes dans le monde chaque année, et elles se manifestent comme des infarctus du myocarde et des insuffisances cardiaques. L'insuffisance cardiaque (IC) survient lorsque le coeur ne peut pas pomper suffisamment du sang pour répondre aux besoins du corps. Les dossiers médicaux électroniques disponibles des patients quantifient les symptômes, les caractéristiques corporelles et les valeurs des tests de laboratoire clinique, qui peuvent être utilisés pour effectuer une analyse biostatistique visant à mettre en évidence des modèles et des corrélations autrement indétectable par les médecins. L'apprentissage automatique, en particulier, peut prédire la survie des patients à partir de leurs données et identifier les caractéristiques les plus importantes parmi celles incluses dans leur dossier médical."

La conclusion des auteurs est la suivante : "Nos résultats de ces modèles à deux caractéristiques montrent non seulement que la créatinine sérique et la fraction d'éjection sont suffisantes pour prédire la survie des patients atteints d'insuffisance cardiaque à partir des dossiers médicaux, mais aussi que l'utilisation de ces deux caractéristiques seules peut conduire à des prédictions plus précises que l'utilisation des caractéristiques de l'ensemble de données d'origine dans son intégralité"

Nous allons vérifier si nous retrouvons cette conclusion: c'est-à-dire nous allons vérifier si le modèle de Cox avec ces deux variables seules est meilleur que d'autres.

Les données contiennent pour 297 patients souffrant d'insuffisance cardiaque sur lesquels on a mesuré

- `age` : age du patient (années)
- `anemie` : diminution des globules rouges ou de l'hémoglobine (boolean)
- `hypertension` : si le patient a de l'hypertension (boolean)
- `CPK` : niveau de l'enzyme CPK dans le sang (mcg/L)
- `diabete` : si le patient a du diabète (boolean)
- `fraction_ejection` : pourcentage de sang quittant le coeur à chaque contraction (percentage)
- `plaquettes` : plaquettes dans le sang (kiloplatelets/mL)
- `sexe` : femme ou homme (binary)
- `creatinine` : taux de créatinine sérique dans le sang (mg/dL)
- `sodium` : taux de sodium sérique dans le sang (mEq/L)
- `fumeur` : si le patient fume ou non (boolean)

- `temps` : temps de suivi (days) (temps censuré)
- `mort` : si le patient est décédé pendant la période de suivi (boolean) (indicatrice de censure)

J'ai créé des nouvelles variables facteurs :

- `EF` à partir de la variable `fraction_ejection` qui prend les niveaux $(14, 30]$, $(30, 45]$ et $(45, 80]$.
- `SC_norm` à partir de `creatinine` qui a les niveaux $(0, 1.5]$, et $(1.5, 10]$.
- `SC_quant` à partir des quartiles de `creatinine` qui a les niveaux $(0.51, 0.9]$, $(0.9, 1.1]$, $(1.1, 1.4]$ et $(1.4, 10]$

1. Représenter l'estimateur de Kaplan-Meier. Quelle est la probabilité pour les patients de survivre 3 mois ou plus ?
2. Tester s'il y a une différence des fonctions de survie pour les 3 groupes de la variable 'EF'. Dans quel groupe de 'EF' les patients ont le plus fort risque de décès ?
3. Faire deux premiers modèles de Cox :
   - l'un avec toutes les variables sauf `creatinine, fraction_ejection` et `SC_quant`
   - l'autre avec toutes les variables sauf `creatinine, fraction_ejection` et `SC_norm`

   Lequel préférez vous ? Grâce à quel critère arrivez vous à cette conclusion ?
4. Faites une sélection de modèle à partir du modèle choisi à la question précédente.
5. Dans le modèle sélectionné, vérifiez par un test si l'âge a un "effet non-linéaire" sur le risque.
6. Vos conclusions sont-elles en accord avec celles de l'article ?
7. Quelle était, au début de l'étude, la probabilité (estimée via le modèle sélectionné) que l'individu 2 du jeu de données soit encore vivant 100 jours après le début de l'étude ? Donnez un intervalle de confiance pour cette prédiction.