TP2 - Le modèle de Cox

Exercice

On étudie la base de données **churn** qui contient des informations sur 7 032 clients d'un grande compagnie. Le but est d'étudier les facteurs de risque pour expliquer les résiliations des clients. La variable d'intérêt est donc le temps écoulé depuis la souscription d'un contrat dans la compagnie jusqu'à la résiliation du client. La variable `tenure` représente la durée observée et la variable `Churn` indique si ce temps est la durée d'intérêt (auquel cas `Churn` vaut "Yes") ou si ce temps est une censure (auquel cas `Churn` vaut "No"). Les covariables observées sont :

- Partner : "Yes" ou "No', indique si le client est partenaire de la compagnie,
- PhoneService : "Yes" ou "No', indique si le client a pris un abonnement téléphonique via la compagnie,
- InternetService : "No" si le client n'a pas pris d'abonnement internet via la compagnie ou "DSL" si le client a pas pris un abonnement DSL ou "Fiber optic" si le client a pas pris un abonnement avec la fibre,
- PaymentMethod : indique le mode de paiement qui peut être "Bank transfer (automatic)", "Credit card (automatic)", "Electronic check" ou "Mailed check",
- Contract : indique le type de contrat souscrit qui peut être "Month-to-month", "One year" ou "Two year",
- PaperlessBilling : "Yes" ou "No', indique si le client reçoit des factures papiers,
- Dependents : "Yes" ou "No', indique si le client a des personnes à charge (par exemple s'il a des enfants).

1. Dans le but de faire une analyse de survie, coder la variable indicatrice de censure `Churn` en 0 et 1. Donner le taux de censure de la base de données.
2. Tracer la courbe de survie globale de la durée d'abonnement estimée par la méthode de Kaplan-Meier. Donner une estimation du premier quartile. Donner une estimation de la probabilité d'avoir résilié son contrat dans la compagnie au bout de 4 ans.
3. Pour chaque variable qualitative, tracer les estimateurs de Kaplan-Meier pour chaque modalité de la variable. Donner également le résultat du test du log-rang pour chaque variable. Commenter chacune des courbes en comparant des estimations de quantiles ou des estimations de survie pour un temps donné. Commenter les résultats des tests du log-rang et faire une liste des variables qui semblent liées à la durée de souscription.
4. Construire un modèle de Cox uniquement avec la variable "InternetService". Donner les rapports de risque pour l'abonnement DSL par rapport à "pas d'abonnement" et pour l'abonnement à la fibre par rapport à "pas d'abonnement". Cette variable a-t-elle un effet significatif sur la durée de souscription ?
5. Rajouter la variable "Contract" au modèle de Cox précédent. Comparer les rapports de risque pour les abonnements internet par rapport à ceux estimés précédemment. Expliquer la différence, on pourra pour cela étudier le lien entre les variables "InternetService" et "Contract".
6. Construire un modèle de Cox avec uniquement la variable "Dependents". Commenter les rapports de risque estimés et le test correspondant. Construire, pas à pas, les modèles suivants :
   - Dependents+Partner

— Dependents+Partner+Contract
— Dependents+Partner+Contract+InternetService

Que se passe-t-il pour l'effet de la variable "Dependents" ? Vous semble-t-il que cette variable a un effet sur la durée de souscription ? Détailler votre réponse.

7. Construire à présent le modèle complet contenant toutes les variables. Commenter l'effet de chaque variable ainsi que le rapport de risque des variables significatives. Donner les intervalles de confiance de ces rapports de risque.
8. À partir du modèle complet de la question précédente, rajouter un terme d'interaction entre les variables "Partner" et "PaperlessBilling". Construire un test statistique permettant de tester si il existe une interaction entre ces deux variables. On pourra pour cela considérer le modèle sans interaction et construire un test du rapport de vraisemblance à l'aide de la commande `anova`. Le test est-il significatif au niveau 10% ? Au niveau 5% ?
9. À partir du modèle complet de la question 7. rajouter un terme d'interaction entre les variables "Partner" et "InternetService". Vous semble-t-il qu'il y a une interaction significative entre ces deux variables ? Calculer le rapport de risque entre les clients étant partenaires et ceux ne l'étant pas parmi les clients qui ont un abonnement internet avec la fibre. Même question pour les clients qui n'ont pas d'abonnement internet et pour ceux qui ont un abonnement avec la DSL.

   Bonus : en utilisant la fonction `estimate` du package `lava` donner les intervalles de confiance de ces rapports de risque avec la p-valeur associée.
10. Proposer une méthode de sélection de variables. Commenter l'algorithme pas à pas. Quelles sont les variables sélectionnées ? On utilisera exclusivement ce modèle dans la suite de l'exercice.
11. À partir du modèle sélectionné à la question précédente, valider le modèle de Cox en utilisant la fonction `cox.zph`. Le modèle de Cox vous semble-t-il valide ? D'après vous, quelle(s) variable(s) ne vérifi(ent) pas l'hypothèse de risque proportionnel ?
12. Construire un modèle de Cox stratifié sur la variable "Contract". Comparer les rapports de risque et les résultats des tests par rapport à ceux obtenus avec le modèle non stratifié.
13. **Bonus** : construire un modèle avec effet dépendant du temps pour la variable "Contract". On veut imposer un rapport de risque pour les temps inférieurs à 25 mois différent du rapport de risque pour les temps supérieurs à 25 mois. Pour cela, on pourra utiliser la `survSplit`. Donner les rapports de risque estimés et conclure sur l'étude de la base de données.