Méthodes d'analyse biostatistique
STT 6510

Projet 2 (A faire en équipes de 2 ou 3).

Date limite de remise le 21 novembre 2023 à 22h par courriel (un seul PDF par équipe, il peut contenir des images ou photos au besoin).
Vaut 15 pourcent de la note totale au cours.

Instructions:

Remettre une seule copie PDF contenant toutes vos réponses (en ordre des questions) par équipe. Pour les questions qui nécessitent du code en R, montrer toutes les parties les plus importantes du code (ex., l'ajustement d'un modèle et la sortie du résumé du modèle). Vous pouvez, entre autres, utiliser Overleaf en ligne pour produire un document que vous pourrez modifier simultanément dans votre équipe.

Question 1. Fonction de survie: Estimation et tests statistiques (total 11 points)

Pour répondre à tous les éléments de cette question, vous utiliserez le jeu de données `diabetic` en R. Il peut être importé en R à partir des commandes `library(survival); data(diabetic)`. Notez qu'un résumé des variables dans ce jeu de données est donné en ligne (entre autres, on peut trouver plus d'information dans le document PDF d'information de la librairie `survival`, en ligne).

a) (1 point) Décrire brièvement toutes les variables disponibles dans le jeu de données `diabetic`, donc donner leur nom et ce qu'elles représentent. De quelle étude ces données proviennent-elles? (c.-à-d., quelle était une question de recherche menant à cette collecte de données?)

b) (1 point) Sous R, produire un graphique avec la courbe de survie Kaplan-Meier pour l'échantillon complet, pour le temps jusqu'à devenir aveugle (défini comme une acuité visuelle tombant sous le 5/200, selon l'information en ligne). Décrire la courbe de survie dans vos mots.

c) (1 point) Stratifier la courbe de survie par groupe traitement (varibale `trt` dans le jeu de données). Ajouter une légende pour indiquer quelle courbe correspond au groupe 0 ou au groupe 1.

d) (3 points) Restreignez-vous maintenant au groupe `trt`= 0, les yeux n'ayant pas reçu de traitement au laser. A l'intérieur de ce groupe, trouvez d'abord le temps de survie médian (c.-à-d., le temps après lequel la probabilité de survie tombe à 50%). Puis, dérivez un intervalle de confiance à 95% pour ce temps de survie médian en utilisant l'approche de Barker (2009) décrite en classe.

e) (2 points) Sous R, faites un test du Log-Rank pour comparer les deux groupes de traitement et discutez du résultat/conclusion du test.

f) (3 points) Comme en e), supposons que l'on s'intéresse à la comparaison des groupes traitement 0 vs 1 en termes de temps jusqu'à devenir aveugle. On se demande si d'autres variables dans le jeu de données pourraient être des facteurs confondants, et si l'on devrait stratifier le test du Log-Rang sur l'une de ces variables. A partir de statistiques descriptives, de graphiques et/ou d'arguments adaptés au contexte de l'étude, discuter de laquelle des variables du jeu de données (autre que les variables traitement, temps de survie, indicateur d'événement) risque d'agir comme facteur confondant, et reproduire le test du Log-Rank stratifié pour cette variable. Qu'observez-vous? Concluez.

Question 2. Modèle de Cox (total 9 points)

Pour répondre à cette question, utilisez le jeu de données `rotterdam` en R. Il peut être importé en R à partir des commandes `library(survival); data(rotterdam)`. Si la dernière option avec data() ne fonctionne pas, essayez tout simplement de travailler sur le jeu de données directement, c.-à-d., utilisez directement `rotterdam` ou faites le `summary(rotterdam)` pour voir si vous y avez accès. Cela fonctionnait mieux de mon côté.

Il s'agit d'un jeu de données sur les temps jusqu'à une récurrence du cancer du sein dans un jeu de données collectées sur des patientes ayant le cancer du sein qui sont en rémission, dans une banque de données de Rotterdam, Pays-Bas. Plusieurs variables sont disponibles dans le jeu de données.

a) (3 points) On s'intéresse à l'association entre des variables explicatives `age`, `meno`, `size`, `grade`, `nodes` et `chemo`, et le temps jusqu'à une **récidive** du cancer du sein. Ajuster un modèle de Cox incorporant toutes ces variables explicatives pour le temps jusqu'à une récidive. Rapporter les rapports de risque (*hazard ratio*) pour chacune des variables explicatives avec leur intervalle de confiance. Interpréter tous ces rapports de risque à partir de phrases expliquant ce que ces rapports de risque signifient dans l'étude, pour ces patientes.

b) (3 points) Pour la question en a), on a fait une hypothèse importante nous permettant d'interpréter le rapport de risque comme le rapport des fonctions de risque en tout temps t, pour différents niveaux des caractéristiques. Quelle est cette hypothèse? Que signifie-t-elle? Sans le faire, discutez de comment vous pourriez évaluer graphiquement cette hypothèse pour le modèle de Cox discuté en a)?

c) (3 points) Décrire dans vos mots les 3 tests statistiques discutés en Sections 5.3.1, 5.3.2, 5.3.3 du livre de Moore (2016) pour les coefficients dans le modèle de Cox. Comme dans la Section 5.3., dans un deuxième temps, discutez/comparez entre les 3 tests leurs avantages et inconvénients. Cette réponse devrait consister en environ 8 à 12 lignes max au total.

Question 3. Lissage pour la fonction de risque (total 10 points)

Pour répondre à cette question, utilisez le jeu de données `datasurv.txt` que j'ai mis sur Studium dans le dossier Projets.

a) (2 points) Pour ce jeu de données, reproduisez le tableau de la diapo. 57 (partie 2 du cours) où vous montrez t_i, n_i, d_i, q_i, $1 - q_i$ et S_i avec une ligne pour chaque temps t_i où un événement (échec) a eu lieu. L'indicateur d'événement est représenté par la variable `event` dans le jeu de données, et le temps jusqu'à un événement ou une censure par la variable `time_days`.

b) (2 points) Produisez un graphique montrant la fonction q_i dans le temps (c.-à-d., q_i en fonction du temps t_i dans le tableau fait en a). Mais, incluez aussi des $q_i = 0$ à tous les temps t où il n'y a pas d'événement, c.-à-d., tous les temps que vous n'avez pas mis dans votre tableau. Considérez une abscisse (axe des X) allant des temps 0 à 500 jours. Cela représente ultimement la fonction de risque estimée dans le temps.

c) (3 points) Estimez une version lisse de la fonction de risque à partir du noyau d'Epanechnikov en faisant les calculs à la main et en n'utilisant pas de logiciel statistique pour obtenir le lissage.

Pour ce faire, suivez plutôt les étapes suivantes:

i) Concentrez-vous seulement sur les temps $t \in 200, 210, 220, 230, 240, 250$ pour calculer l'estimateur lisse en ces points seulement.

ii) Estimez, pour chacun des points ci-hauts, la fonction $\hat{h}(t)$ vue en diapositive 74 (partie 2 du cours) en utilisant un paramètre de lissage égal à $b = 5$. Rapportez les 6 valeurs trouvées.

d) (3 points) Comparez les estimateurs pour la fonction de risque suivants dans deux graphiques séparés, mis côte-à-côte:

i) un graphique de la fonction $\hat{h}(t) = d_i/n_i$ (et 0 aux temps où il n'y a pas d'événement) entre les temps $t \in [200, 260]$, basé sur les sous-questions a-b), et

ii) un graphique de la fonction lissée $\hat{h}(t)$ à partir du noyau d'Epanechnikov obtenue au point c) ci-haut où vous reliez les estimations $\hat{h}(t)$ correspondant aux points 200, 210,..., 260 par des lignes droites, avec l'axe des X allant de $t \in [200, 260]$.

Discutez des résultats de cette comparaison.