Méthodes d'analyse biostatistique projet 2 (date limite : le 21 novembre)

Wen, Zehai; Li, Qingyue

2023-11-11 17:44:57-05:00

Exercice 1

Décrivez brièvement les variables du jeu de données `diabetic` dans la bibliothèque `survival` en R. Quelle était une question de recherche menant à cette collecte de données? Produisez un graphique de Kaplan-Meier utilisant l'échantillon entier, le groupe de `trt = 0` et le groupe de `trt = 1` pour le temps jusqu'à devenir aveugle. Pour le groupe `trt = 0`, trouvez le temps médian de survie et construisez un intervalle de confiance à 95% pour le temps médian de survie. Faites un test de Log-Rang pour comparer les deux groupes.

Finalement, 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. À 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 risque d'agir comme facteur confondant, et reproduire le test du Log-Rank stratifié pour cette variable.

Solution à faire

```
quelque listing
```

Exercice 2ab

Pour le jeu de données `rotterdam` dans la bibliothèque `survival` en R, on s'intéresse par l'association entre les variables explicatives `age`, `meno`, `sizes`, `grade`, `notes` et `chemo`, et la variable de réponse le temps jusqu'à une récidive du cancer au sein. Ajustez un modèle de Cox pour cela et rapporter les rapports de risque pour toutes les variables explicatives avec leurs intervalles de confiance (à 95% ?). Nommez l'hypothèse importante du modèle de Cox. Expliquez sans le faire comment on peut vérifier cette hypothèse.

Solution à faire

Exercice 2c

Décrivez dans vos mots les trois tests statistiques discutés en Sections 5.3.1, 5.3.2, 5.3.3 du livre "Applied Survival Analysis Using R" 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 trois tests leurs avantages et inconvénients. Cette réponse devrait consister en environ 8 à 12 lignes max au total.

Solution Écrire $h_i(t_j)$ pour le hasard de patient i au temps d'échec t_j. Si l'hypothèse de proportionnalité est vraie, on peut écrire $h_i(t_j) = h_0(t_j)\psi_i = h_0(t_j)e^{z_i\beta}$ où h_0 est le hasard de référence, $z_i \in \{0, 1\}$ selon si le patient i est dans le groupe de traitement ou le groupe de contrôle et β est le paramètre à estimer. La vraisemblance partielle est définie par $L(\beta) = \prod_{k=1}^{D} \frac{\psi_i}{\sum_{l \in R_k} \psi_l}$ où D est le nombre d'échecs et R_k est l'ensemble des patients à risque au temps t_k. On teste l'hypothèse nulle $H_0 : \beta = 0$. Fixer un niveau $\alpha \in (0, \frac{1}{2})$. Écrire $l = \ln(L)$, $S = l'$, $I = -l''$ et soit $\hat{\beta} \in \arg\max_\beta L(\beta)$. Le test de Wald rejette H_0 si $\left|\hat{\beta}\sqrt{I(\hat{\beta})}\right| > z_{\frac{\alpha}{2}}$. Le test du score rejette H_0 si $\left|\frac{S(0)}{\sqrt{I(0)}}\right| > z_{\frac{\alpha}{2}}$. Le test du ratio vraisemblance rejette H_0 si $\left|2(l(\hat{\beta}) - l(0))\right| > \chi^2_{\frac{\alpha}{2},1}$. Si on ne peut pas obtenir $\hat{\beta}$, il faut utiliser le test du score. Supposons à partir de maintenant qu'on a obtient $\hat{\beta}$. Le résultat du test du ratio vraisemblance a un avantage par rapport aux deux autres tests qu'il est invariant contre les transformations monotones. Si l'invariance n'est pas nécessaire, alors le test de Wald est plus commun et plus facile à calculer. ////

Exercice 3

Pour le jeu de données `datasurv.txt`, écrire $t_1 < \cdots < t_D$ pour tous les temps distincts d'échec, n_i pour le nombre de sujets à risque au temps t_i et d_i pour le nombre d'échecs au temps t_i. Comme les patients à risque inclut ceux qui ont eu un échec, on a $n_i \geqslant d_i \geqslant 1$ pour tout i. On définit $q_i = \frac{d_i}{n_i}$ et $S(t) = \prod_{t_i \leqslant t}(1 - q_i)$ pour tout $t \in \{1, \cdots, t_D\}$. Produisez un tableau de valeurs de t_i, n_i, d_i, q_i et $S(t_i)$.

En R, réalisez un graphique de q en fonction de t pour $t \in \{1, \cdots, 500\}$. Ici, $q(t) = q_i$ si $t \in \{1, \cdots, t_D\}$ et $q(t) = 0$ sinon. Produisez en particulière le même graphique pour $t \in \{200, \cdots, 260\}$ Finalement, lissez cela utilisant le noyau d'Epanechnikov avec un paramètre de lissage 5. Plus précisément, on définit le noyau d'Epanechnikov comme $K(u) = \frac{3}{4}(1 - u^2)\chi_{[-1,1]}(u)$ pour tout $u \in \mathbb{R}$ et on définit la fonction de lissage comme :

$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^{D} K\left(\frac{t - t_i}{b}\right) q_i$$

Ici $b = 5$. Produisez un graphique de $\hat{h}$ en fonction de t pour $t \in \{200, \cdots, 260\}$. Vérifiez les résultats en main pour $t = \{200, 210, 220, 230, 240, 250\}$.

Solution à complet