

STT 3260

Examen 2 à emporter: Projet - Automne 2018

- **Vous ne pouvez pas commenter ou parler de cet examen avec personne sauf les membres de votre groupe, le démonstrateur et le professeur.**
- **Toute sorte de collaboration est fortement interdite.** Si on trouve que votre examen est suspect de collaboration avec un autre groupe ou un membre d'autre groupe, il se peut que vous et les membres de votre groupe seront demandés de parler avec le professeur pour fixer une date pour un examen orale (la date et l'heure de l'examen seront fixés par l'instructeur).
- Pour recevoir tous les points dans chacun des problèmes vous devez clairement expliquer comment vous êtes arrivé à vos réponses.
- L'examen doit être retourné au professeur le mardi 4 décembre 2018 avant 17h00.

Indications

Chaque groupe doit écrire un rapport de pas plus de huit pages (cinq pages plus trois pages de code) avec une taille de police (caractère) minimum de 12pt. Le rapport doit être divisé en deux parties:

1. **Un résumé technique, de pas plus de cinq pages,** contenant les détails de votre analyse statistique. L'idée est de décrire votre analyse à un public instruit dans la statistique dans une façon claire et organisée. Aucun appendice/annexe (autre que le code utilisé) sera admis.

La clarté, la concision et l'organisation dans le rapport sont des caractéristiques importantes. Elles seront appréciées et récompensées. Il est très important que vous soyez capable de communiquer ce que vous avez fait à des autres personnes.

En écrivant le rapport considérez les trois parties suivantes: introduction, méthodes et résultats, discussion et conclusions.

2. **Une appendice avec le code SAS ou/et R** que vous avez utilisé pour résoudre les problèmes (n'utilisez pas plus de trois pages pour l'appendice).

Objectifs

- Appliquer les notions théoriques vues en classe sur un problème pratique.
- Résoudre un problème important d'analyse de survie, dans son ensemble et non seulement en tant qu'éléments fragmentaires.
- Améliorer vos connaissances du logiciel SAS et/ou R. SAS est fortement utilisé en statistique et en actuariat.

Il se peut que même si vous travaillez durement vous n'obtiendrez pas une réponse satisfaisante au problème. La façon dans laquelle vous abordez le problème est plus importante que la réponse à laquelle vous arrivez à la fin. Quelques fois les données peuvent admettre plus d'une réponse satisfaisante; quelques fois il n'y a pas de réponse satisfaisante. S'il se passe que vous n'arrivez pas à une réponse satisfaisante pour vous même, SVP mentionnez des autres façons dans lesquelles vous avez abordé le problème et les raisons pour lesquelles vous les avez abandonnées. Si vous arrivez à plusieurs réponses que vous pensiez correctes, mentionnez le pour et le contre associé à chacune des réponses.

Enfin, rappelez vous que l'analyse de données est un outil pour essayer de comprendre un phénomène ou de résoudre un problème scientifique. Essayez donc de mettre l'analyse statistique dans le contexte dans lequel les données ont surgi.

Le Problème et les données

Les données pour ce projet proviennent d'un essai clinique comprenant des individus atteints de fibrose kystique. Ces personnes sont sensibles à une accumulation de mucus dans les poumons, ce qui conduit à l'exacerbation pulmonaire et à la détérioration de la fonction pulmonaire.

Le but de l'étude est d'évaluer l'efficacité de l'administration quotidienne d'une forme recombinante de l'enzyme humaine DNase 1 dans la prévention des exacerbations. Les individus ont été attribués au hasard, soit le nouveau traitement (appelé *Pulmozyme*), soit un placebo. Il y a 324 individus dans le groupe placebo, et 321 individus dans le groupe Pulmozyme.

Les individus, qui ne subissaient pas des exacerbations avant la randomisation, ont été suivis pendant environ 169 jours. Les jours au cours desquels une période d'exacerbation avait commencé ont été enregistrés. Considérez comme variable d'intérêt T = le temps à la première période d'exacerbation ou infection.

Quand une période d'exacerbation avait commencé, l'individu a été administré des antibiotiques, et après que l'exacerbation avait disparu, l'individu était alors considéré comme à risque d'une nouvelle exacerbation. Par conséquent, certains individus n'ont pas eu des exacerbations au cours des 169 jours de la période de suivi, certains en ont eu un, et certains en ont eu deux ou plus.

Une autre variable concomitante et continue, le volume expiratoire forcé (**vef**) a également été mesurée sur chaque individu au début de l'étude. Cette variable est une mesure de la fonction pulmonaire, et on pourrait s'attendre à ce que des grandes valeurs du **vef** soient associées à des durées de survie plus grandes.

Analysez ces données en utilisant la méthodologie d'analyse de survie pour comprendre la relation entre le temps de survie et les traitements, et le rôle que **vef** joue dans les temps de survie. Expliquez et interprétez vos résultats.

En rédigeant votre rapport, prenez en considération les points et questions suivantes:

1. Explorer les données. Faites une description de toutes les variables de l'étude (sauf la variable "temps de survie"); déterminez la moyenne et l'écart-type pour les variables continues, les fréquences pour les variables dichotomiques.
2. Vérifiez si l'étude était randomisée par les valeurs de **vef**, c'est-à-dire, réalisez un test non-paramétrique pour vérifier si les distributions de **vef** dans les deux groupes sont égales.
3. Faites l'estimation non-paramétrique des fonctions de survie par traitement. Est-ce qu'il y a des différences entre les taux de risques des traitements?
4. Pour chaque groupe de traitement, divisez la variable **vef** en trois groupes: valeurs bas (**vefB**), valeurs médianes (**vefM**), et valeurs élevées (**vefE**). Est-ce qu'il y a des différences entre les taux de risques associés à ces six groupes? Si nécessaire, faites des comparaisons simultanées entre tous les paires possible de ces six groupes. Regardez aussi les six estimateurs de courbes de survie et commentez.
5. Trouvez aussi des intervalles de confiance pour chacune des six fonctions de survie évaluées aux centiles (percentiles) 10% et 30%. Commentez.
6. Ajustez un modèle de régression de défaillance accélérée paramétrique aux données (par exemple, de Weibull, log-normale, log-logistique, etc.). Faites un test approprié pour vérifier si le modèle est adéquat. Vérifiez aussi si le modèle est approprié en utilisant l'analyse des résidus (diagnostic du modèle).

Remarque: Vous devez voir si l'incorporation de la variable continue **vef** explique mieux les temps de survie observés. Faites le même exercice avec les variables **vefB**, **vefM**, **vefE** au lieu de **vef**. Quel modèle choisiriez vous?

7. Ajustez un modèle de régression de Cox aux données. Est-ce que l'hypothèse des risques proportionnelles semble correct ? En ce qui concerne l'effet de la variable traitement, Est-ce que l'effet change avec le temps de survie ? Considérez l'incorporation des variables d'interaction $X = \text{traitement} \times \mathbf{I}\{t \geq \tau\}$ et $Z = \text{vef} \times \mathbf{I}\{t \geq \tau\}$ au modèle de Cox, où τ est une valeur adéquate fixée par vous, et

$$\mathbf{I}\{t \geq \tau\} = \begin{cases} 1, & \text{si } t \geq \tau, \\ 0, & \text{si } t < \tau. \end{cases}$$

8. Comparer le modèle de régression de défaillance accélérée avec le modèle de régression de Cox.
9. Quelles sont vos conclusions?

Les données se trouvent sur le site studiUM du cours. La documentation de SAS 9.1.3 est disponible en ligne à l'adresse suivante :

https://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/stat_ug_7313.pdf