Méthodes d'analyse biostatistique
STT 6510

Projet 1 (A faire en équipes de 2 ou 3).

Date limite de remise le 17 octobre 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. J'encourage fortement les étudiant.e.s à apprendre LaTex le plus rapidement possible, vous en aurez besoin tôt ou tard!

Question 1. Théorie, 5 sous-questions a) à e) pour un total de 10 points.

a. (2 points) Montrer que la loi de Poisson fait partie de la famille exponentielle et montrer les fonctions $a(\phi)$, $b(\theta)$ et $c(y, \phi)$ correspondantes pour la famille exponentielle

b. (2 points) Supposons qu'on ajuste un modèle de régression logistique pour une variable binaire Y en fonction de variables explicatives X_1, X_2 et X_3, où X_1 est une variable binaire prenant les catégories 0 et 1, X_2 est binaire, et X_3 est une variable continue.

i) Montrez que pour des valeurs fixes (c.-à-d. constantes) de X_2 et X_3, le coefficient pour X_1 dans la régression peut être interprété comme un log-rapport de cotes.

ii) Dans un deuxième temps, trouvez une expression pour le log-rapport de cotes pour deux personnes comparées: une ayant $X_1 = 1$ et $X_3 = 7$, et une autre personne ayant $X_1 = 0$ et $X_3 = 5$, si les deux ont la même valeur pour X_2.

c. (2 points) Démontrez le résultat pour la variance présenté à la diapositive 40 dans les notes de cours de la partie 1 (c'est-à-dire, montrer toutes les étapes pour arriver à $Var(Y \mid X) = \exp(\beta_0 + \beta_1 X)\alpha$ incluant l'expression pour α, montrant ainsi la sur-dispersion).

d. (1 point) Expliquez assez brièvement, pourquoi les modèles de régression linéaire généralisés (c.-à-d. les GLM) ne nécessitent pas réellement la spécification complète de la distribution de Y pour l'estimation des coefficients.

e. (3 points) Soit le modèle mixte suivant:

$$Y_{ij} = a_i + b_i X_{ij} + \beta X_{ij} + \epsilon_{ij}$$

où $i = 1, ..., m$, $j = 1, ..., 5$ et on assume

$$\begin{aligned} a_i \sim & \mathcal{N}(4, \tau^2) \\ b_i \sim & \mathcal{N}(0, \psi^2) \\ Cov(\epsilon_{ij}) = & \sigma^2 V \\ V = & \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \end{aligned}$$

c'est-à-dire que les erreurs ont une corrélation de type $AR(1)$ entre elles pour un même individu i. Les effets a_i, b_i dépendent du patient et sont des effets aléatoires (ordonnée et pente). On assume que la matrice X_{ij} ne contient qu'une seule colonne, c'est-à-dire que le modèle inclut une seule variable explicative plus l'ordonnée à l'origine.

i) Séparez et présentez le modèle comme deux composantes: une composante de régression linéaire et une erreur W_{ij}, similairement à ce qui a été fait en classe.

ii) Présentez les éléments de la matrice de $Cov(Y_i)$, c'est-à-dire, la variance (éléments de la diagonale) et la covariance entre les mesures d'un même individu i sous le modèle hiérarchique ci-haut.

Question 2. Questions variées en lien avec des analyses de données, total de 11 points. A faire avec le logiciel R.

a. (2 points) Reprendre le jeu de données sur les limules disponible à partir du progiciel `rsq` en R. Ajustez un modèle de régression de Poisson pour le nombre de satellites mâles en fonction des variables explicatives suivantes: *color*, *spine*, et *width*. Dans un deuxième temps, évaluez la sur-dispersion potentielle en faisant un test de sur-dispersion et en évaluant le paramètre de sur-dispersion. Discuter des résultats.

b. (1 point) Supposons que le test de sur-dispersion en 2a) rejète l'hypothèse nulle. Expliquez ce que rejetter cette hypothèse signifie, et suggérez une approche alternative, sans conduire l'analyse alternative.

c. (1 point) Utilisez un test statistique approprié (sous R) pour comparer toutes les déclinaisons possibles du modèle en a), en n'assumant pas de sur-dispersion (donc un modèle de Poisson de GLM classique): c'est-à-dire, comparez les modèles sous toutes les combinaisons possibles des variables explicatives *color*, *spine*, *width*, et *weight*. Garder le meilleur modèle selon vos tests (et le/les critères utilisés), le présenter.

d. (2 points) Pour le modèle choisi au point c) précédent, présenter une analyse des résidus et discuter ce que vous observez. Vous pouvez choisir le types de résidus que vous souhaitez.

e. (5 points) Utilisez le progiciel `geex` sous R (ou toute autre fonction permettant de maximiser ou optimiser sous R) pour estimer les paramètres du modèle suivant avec une approche d'équation d'estimation de votre choix (tout en discutant des hypothèses que vous faites, si vous en faites):

$$E[Y_i \mid x_i] = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Ultimement, le modèle sera utilisé pour estimer dans le jeu de données `addhealth` mis sur Studium l'association entre la variable `weight` et les variables explicatives `age` et `SES` (statut socioéconomique). Une description du jeu de données est disponible plus bas. C'est-à-dire, pour

répondre à cette question, vous devez **suivre les étapes suivantes**:

1) Postuler une fonction d'équation qui permet d'estimer de façon convergente les coefficients dans le modèle pour Y ci-haut, montrer l'équation choisie et discuter de pourquoi vous avez choisi cette équation dans votre devoir. Montrer le code R utilisé pour définir l'équation d'estimation

2) Importer le jeu de données `addhealth` sur R

3) Appliquer la méthode d'estimation au jeu de données `addhealth` et obtenir des estimés pour les paramètres $\beta_0, \beta_1, \beta_2$ dans le modèle

$$E[\text{Weight}_i \mid \text{age}_i, \text{SES}_i] = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{SES}_i,$$

puis décrire dans vos mots (sans les calculer) comment vous pourriez estimer la variance des coefficients estimés.

Je donnerai aussi un **Point Bonus** d'1 point pour ceux qui calculeront (obtiendront) les estimés de la variance des coefficients β, s'ils sont calculés convenablement, mais vous n'êtes pas obligés de les calculer pour avoir vos points en question e).

(Truc: Un article discuté en classe sur la fonction `geex` pourrait s'avérer utile en cas de doute.)

Description du jeu de données `addhealth`:

Le jeu de données consiste en les données de la première vague de l'étude longitudinale Add Health qui porte sur la santé des adolescents aux États-Unis (time=1 dans le jeu de données, car cela correspond à la première vague seulement). Des détails sur l'étude se trouvent ici : `https://addhealth.cpc.unc.edu/` ainsi qu'ici: `https://core.ac.uk/download/pdf/210590553.pdf`. En gros, le jeu de données disponible pour votre analyse (qui consiste seulement en une sélection des données complètes de l'étude) contient des informations sur le sentiment d'être déprimé, le statut de fumeur, le poids, l'âge, le sexe, et le statut socioéconomique des adolescents durant la première vague de l'étude (ainsi qu'un identifiant pour chaque individu).

Question 3. (9 points) Analyse de données longitudinales. Cette question utilise le jeu de données longitudinales *addhealth long* qui contient des observations répétées pour chaque individu dans le temps. On s'intéresse à l'ensemble de variables suivantes: `age`, `sex`, `SES`, `smoking` et `weight`, notant que les variables `ID` et `time discrete` représentent respectivement l'identifiant de l'individu et le temps (il y a quatre temps: vague 1, 2, 3 ou 4) dans le jeu de données.

i) (2 points) Selon vous, quelle structure de corrélation pourrait être présente dans ces données, entre les mesures de poids de chaque individu? Discutez avec des arguments (plutôt subjectifs).

ii) (3 points) Vérifiez si les variables explicatives sont importantes pour expliquer la corrélation entre les mesures répétées des individus dans cette étude: Pour ce faire, comparez les trois modèles suivants en termes de la corrélation résiduelle (c.-à-d. la corrélation des résidus résultant de ces modèles):
1) un modèle de régression linéaire pour la variable dépendante `weight` (le poids en livres) en fonction d'aucune variable explicative (c.-à-d. juste une ordonnée à l'origine),
2) un deuxième modèle pour le poids en fonction des variables explicatives `age` et `sex`,
3) puis un troisième modèle dans lequel vous incorporez les variables `smoking`, `SES`, `age` et `sex`, comme variables explicatives du poids (`weight`).
Que trouvez-vous?

iii) (4 points) Supposons que l'on s'intéresse particulièrement à **l'association entre le poids**

et le statut socioéconomique. Comparez les estimés ponctuels (et estimés de variance du coefficient) obtenus à partir des approches suivantes:
1) Un modèle linéaire ordinaire (OLS) où on assume l'indépendance entre toutes les observations, ajusté à partir de la fonction `lm` en R
2) Un modèle linéaire général où on assume une matrice de corrélation interchangeable à l'intérieur d'un individu, pour la réponse `weight` (notre Y)
3) Un modèle GEE où on assume une matrice de corrélation interchangeable à l'intérieur d'un individu, pour la réponse `weight` (notre Y)
4) Un modèle mixte dans lequel vous introduisez une ordonnée à l'origine et une pente aléatoires pour chaque individu et des erreurs indépendantes.

Comparez les modèles à partir d'un critère de votre choix et discutez des résultats. Quel estimé fourniriez-vous (avec son intervalle de confiance) à un client qui se demande quelle est l'association entre ces deux variables? Pourquoi avoir choisi celui-ci? (conseil: pensez aux hypothèses faites sous chaque modèle)