Numéro d'étudiant : …….

Nom : ……

Prénom : ……

Université de Bordeaux - UF Mathématiques et Interactions

M2 Modélisation statistique et stochastique (4TSS901S, S9)
M2 CMI Statistique et informatique (4TST901S, S9)

« Statistique et analyse de données en grande dimension »

CONTROLE CONTINU

27/10/2022

14h-17h

Enseignante : Marta Avalos-Fernandez

Déroulement du devoir :
- Devoir individuel. Vous n'avez pas le droit de communiquer entre vous.
- Vous pouvez utiliser votre ordinateur portable ou un ordinateur de la salle.
- Vous avez droit à tous vos documents et codes, ainsi que à internet. Vous pouvez utiliser un traducteur anglais/français pour l'article, si vous en avez besoin. Le lien vers l'article associé au sujet vous a été envoyé par mail le 19/10/2022.
- Vous pouvez répondre aux questions en anglais ou en français, à la suite des questions (l'espace laissé pour les réponses n'est pas proportionnel à la longueur de réponse attendue) ou sur un document vierge (en précisant le numéro de la question) ou sur papier libre (que vous devrez ensuite scanner avec une application sur téléphone portable). Le fichier doit être transformé au format pdf avant l'envoi, si plusieurs documents ont été créés, ils devront être fusionnés avant l'envoi.
- L'ensemble de réponses de la Partie I sont à envoyer par mail à marta.avalos-fernandez@inria.fr dans un seul fichier pdf : **devoir271022_ nometudiant.pdf avant 15h15.**
- Vous recevrez pour 15h la Partie II du devoir. Vous pouvez y travailler dès que vous avez envoyé votre Partie I.

Partie I

Ce devoir s'appuie sur des extraits de l'article Shah et al. Blood-Based Fingerprint of Cardiorespiratory Fitness and Long-Term Health Outcomes in Young Adulthood. *Journal of the American Heart Association*, 2022;11:e026670.

La forme cardio-vasculaire (CRF) est un bon prédicteur de l'état de santé. La CRF est mesurée par le pic de la consommation d'oxygène (peak VO2) à l'aide d'un test d'effort cardiopulmonaire. Cette mesure est difficile d'obtenir en routine et donc la CRF est sous-utilisée en prévention primaire. La CRF devrait toutefois pouvoir être prédite sans l'utilisation d'un test d'exercice à partir de facteurs de risque cliniques et de biomarqueurs (ou métabolites) sanguins. L'objectif du travail décrit dans l'article était d'obtenir une bonne prédiction du pic de la consommation d'oxygène à partir de facteurs de risque démographiques et cliniques (âge, sexe, indice de masse corporelle, pression artérielle systolique, statut tabagique, cholestérol total, diabètes, antécédent de maladie cardio-vasculaire, ...) et de métabolites sanguins en utilisant les données Framingham Heart Study (FHS). A partir de cette prédiction, les auteurs ont construit un score et l'ont évalué sur les données d'une autre étude.

L'objectif du devoir est de repérer dans vos cours les éléments qui permettent de comprendre la démarche des auteurs (pourquoi, comment) et d'avoir un regard critique vis-à-vis de cette démarche, de leurs résultats et de leurs conclusions et cela sans forcément avoir des connaissances du domaine d'application.

MÉTHODES-STATISTICAL ANALYSIS-DEVELOPMENT OF A METABOLOMIC FITNESS SCORE IN THE FHS

1) Il est indiqué : "LASSO regression was selected to
   - avoid overfitting,
   - address collinearity across metabolites, and
   - provide a parsimonious multimetabolite model for peak VO2. "

Expliquez avez vos propres mots ces arguments pour utiliser la méthode Lasso.

Pour éviter le sur-ajustement il va supprimer les variables les moins pertinentes, pour s'attaquer à la corrélation aux métabolites (ici les prédicteurs), étant donné qu'ils sont corrélés, une co-linéarité importante -> rupture de dimension ou suppression de variables et gérer la corrélation entre les variables explicatives. Si on a une corrélation très forte entre 2 variables, le LASSO va mieux gérer les 2 variables sans faire exploser la variance, il peut en supprimer une ou garder les 2 mais dans tous les cas il va pas foirer. Il est + utile qu'une méthode classique.
Le dernier point c'est pareil que le 1er : il va supprimer des variables qui sont censées être moins importantes donc on va obtenir un modèle plus économe, plus parcimonieux pour expliquer le pic de VO2.

2) Il est indiqué : "Our aim was to develop a metabolite score independent of standard risk factors, so we forced (unpenalized) adjustment for age, sex, body mass index (BMI), systolic blood pressure, hypertension treatment status, total cholesterol, high-density lipoprotein cholesterol, prevalent CVD, diabetes, and smoking status in the LASSO model."

Expliquez ce paragraphe (pourquoi et comment).

Je veux savoir quelles métabolites sont importantes, je ne veux pas que ce qui est du à l'age soit ignoré. je veux éviter les problèmes de confusion, seulement ce qui est proche aux métabolites. On les force à être dans le modèle, sur R on va utiliser : penalty.factor=…
je ne veux pas que ces variables soient touchées, je sais déjà qu'elles ont un impact, elles ont un impact, on ne regarde que les métabolites
Métabolites liées à l'âge, c'est l'age qui explique des choses, je ne veux qu'il vienne prendre le rôle de l'age, si je mets l'age et si je le force à être la je ne vais pas laisser rentrer cette métabolite. Pour les variables forcées dans le modèle je mets 0

Obliger ces variables à être dans le modèle
avec Lasso on le fait avec la penalty.factor (on veut pas que la variable soit touché et se concentrer sur les métabolites)
on veut pas que âge prend tout, on veut laisser la place aux métabolites

Celles ajoutées sont celles qui vont optimiser la prédiction.

3) Il est indiqué : "Continuous covariates and peak VO2 were log-transformed; forced adjustments in LASSO were not standardized."
Pourquoi, à votre avis,
- les variables explicatives continues ont été log-transformées ?
- la variable réponse, peak VO2, a été log-transformée ?
- les variables d'ajustement ne sont pas standardisées ?

Est-ce que, à votre avis, les métabolites sont standardisés ? Pourquoi ?

Pour la normaliser. Dans la régression linéaire, on suppose que les résidus sont gaussiens, centrées, et les variances qui sont les mêmes, des obsevations qui sont iid. Si j'ai une réponse qui est conditionnée par rapport aux variables explicatives pas '...'
Quand on fait des transformations sur la variable réponse, dans les modèles linéaires, on cherche la normalité ou homoscédasticité. Mais c'est pas correcte de le dire pour la variable explicative. Variables réponse + explicatives pas gaussienne = pas valide On a uniquement besoin de la normlaté de la réponse. (il faut pas que ce soit l'unité qui impacte le choix du modèle)
Un argument : « standardized » par défaut à vrai. Les beta est impacté par les unités (minutes, heures, beta X 60) c'est l'opération pour passer d'une unité à une autre.

Si j'ai un beta énorme car l'unité est très petite, il va être comme si il avait une grande importance dans le modele donc il faut systématiquement standardiser les variables quantitatives. Le sexe etc est qualitatif et y'a pas d'unité donc elles ont pas besoin d'être standardisées car variables qualitatives et dépendent pas d'une unité.
Etant donnée que les variables sont pénalisées, elles ne rentrent pas dans la fonction : qu'elles soient standardisées ou non, elles auront aucun impact.
2 raisons: soient pcq les variables sont quali (pas vraies pour tout par ex pour l'age ou la masse corporelle), soient qu'elles ont une pénalité (standardisées ou non)

4) Détaillez le modèle qui a été utilisé par les auteurs ainsi que les hypothèses sur lesquelles repose ce modèle.
Donnez ensuite le problème d'optimisation qui est résolu en utilisant la méthode Lasso (tenez en compte les choix effectués par les auteurs mentionnés dans les questions précédentes).
Soyez rigoureux dans les notations mathématiques.

Pour les variables d'ajustement, faire attention si c'est binaire : qu'un seul beta, si c'est à plusieurs modalités : plusieurs variables.
L'indépendance des observations, la linéarité, la normalité (faire des phrases). Regarder le probleme qu'on fait quand on applique le LASSO.

RÉSULTATS- DEVELOPMENT AND CHARACTERIZATION OF A METABOLOMIC FITNESS SCORE IN THE FHS

5) La figure 2A montre les coefficients de régression (d'une part) des métabolites estimés avec la méthode Lasso.
Combien de métabolites (nombre total de variables explicatives mesurées, moins les facteurs de risque démographiques et cliniques) pourraient être représentés ?
A votre avis, pourquoi seulement ~70 métabolites sont présentés dans la figure 2A ?

Au départ, il y avait 200 métabolites mais environ 70 sont présentés dans la figure 2. Ceux qui sont présentés sont uniquement ceux qui sont différents de 0, qui sont significatifs. On pourrait avoir jusqu'à 200, mais il ya uniquement 70, car ce sont uniquement ceux qui n'ont pas été supprimés dans le modèle, ceux qui ne sont pas nuls.

6) Il n'est pas précisé comment ces estimations de la figure 2A sont obtenues.
Indiquez une méthode avec laquelle ces estimations sont obtenues usuellement. Expliquez brièvement le raisonnement de cette méthode.

Une méthode qui aurait pu le permettre est : la validation croisée (écart type ou minimum de réduction moyen).
Expliquer la méthode.
Méthode analytique AIC

7) Les auteurs ont utilisé une partie de la cohorte FHS pour estimer les coefficients de régression (*derivation subsample* N=451) et la partie restante (*validation subsample* N=914) pour évaluer le coefficient de corrélation entre la partie du modèle correspondant aux métabolites (estimée à partir de la *derivation subsample*, soit le score que les auteurs développent) et la réponse.

Pourquoi, à votre avis, les auteurs ont suivi cette démarche ? Pourquoi, à votre avis, le coefficient de corrélation obtenu avec la *derivation subsamble* est meilleur que celui obtenu avec la *validation subsample* (figure 2B) ?

8) La table 2 montre les résultats du modèle linéaire estimé en utilisant la méthode des moindres carrés (sans sélection de variables) à partir de la *validation subsample*. Onze variables sont utilisées : 10 variables d'ajustement et le score des métabolites (dont les coefficients ont été estimés à partir de la *derivation subsample*).
Donnez l'interprétation de $\hat{\beta}$ associé à *Smoking* ($\hat{\beta}$=−0.193), sachant que *Smoking*=1 si fumeur ou ancien fumeur et *Smoking*=0 pour ceux qui n'ont jamais fumé, et celui associé à *Metabolomic fitness score* ($\hat{\beta}$=0.303).
Pour cela,
- Soit vous tenez en compte dans votre interprétation de la transformation de la réponse.
- Soit si vous préférez (car plus facile) supposez que la réponse peak VO2 n'a pas été transformée, et qu'elle est mesurée en mL/kg per min.

Pour interpréter beta, en moyenne, les fumeurs ont 0.2 unité mL/kg de peak VO2 inférieur aux non fumeurs.
L'important est de savoir que pour une unité d'augmentation de score, en moyenne la variable réponse est augmentée de 0.3.