H. Charles – 4BIM ML & Stat NP
Décembre 2020

Modèle linéaire et statistiques non-paramétriques
4-BiM – 2020
Tous documents autorisés

***Compétences évaluées** : **C2**.Optimiser des plans d'expériences ; **C7**. Choisir et mettre en œuvre des outils statistiques appropriés ; **C8**. Apprécier les limites de validité d'un modèle et identifier les sources de variabilité et d'incertitudes ; **C9**. Modéliser et interpréter des données biologiques pour comprendre les processus sous-jacents.*
***Connaissances et capacités évaluées** : tests d'hypothèses, modèles linéaires mixtes, statistiques non paramétriques, maitrise de R, transposer un problème biologique en un modèle statistique approprié, interpréter un résultat statistique relatif à une question biologique.*

*Le jeu de données[1] (**goodlife.txt**) est dans le répertoire « Jeux de données » du cours Moodle BS-BS-4-BMSTAT5-S1. Les données numériques (émissions de CO2 et consommation d'eau) sont normalisées par habitant et par équivalent planétaire, ainsi une valeur de 1 correspond à un équilibre théorique de la ressource considérée. Les variables qualitatives (continent, region_monde, pays, état sanitaire et niveau de démocratie) sont directement compréhensibles. Attention, le séparateur décimal est une « , ».*

Barème sur 22 points**. **Attention, les 2 dernières questions (Q 12 et 13) comptent pour 5 points. Elles sont indépendantes et peuvent être réalisées sans avoir fait les autres questions.

Dans cet article, les auteurs ont recueilli des données quantitatives sur les empreintes écologiques de différents pays du monde, ainsi que des données sociales. L'objectif de ce travail va être d'analyser ces données en suivant les questions ci-dessous.

Q1 (1 pt). 1. Les variables « sanitaire » et « democratie » sont-elles dépendantes ? Justifier votre réponse par un test approprié et exprimer la conclusion avec une phrase complète.

On souhaite maintenant s'intéresser aux émissions de CO2 en fonction de l'état sanitaire des différents pays et de leur niveau de démocratie. La variable pays ne sera pas utilisée dans cette étude.

Q2 (1 pt). Décrire le plan d'expérience correspondant à l'analyse demandée.

Q3 (2pt). Comparer les émissions de CO2 des pays en fonction de leur niveau de démocratie et de leur état sanitaire par un modèle (lm1) et réaliser les tests appropriés ?

Q4 (1 pt). Ecrire le modèle lm1 sous la forme d'une équation et interpréter les termes significatifs.

[1] ***:** O'Neill, D.W., Fanning, A.L., Lamb, W.F., and Steinberger, J.K. (2018). A good life for all within planetary boundaries. Nature Sustainability 1, 88-95. doi: 10.1038/s41893-018-0021-4 (les données ont été simplifiées et formatées pour l'exercice).*

Q5 (1 pt). Changer l'ordre d'introduction des variables explicatives de votre modèle ? Quels effets observez-vous sur votre analyse (lm2) et pourquoi ?

Q6 (2 pt). Proposer une solution pour avoir des estimations correctes de votre décomposition de la variance et refaire les conclusions de l'analyse.

Q7 (2pt). Vérifier les hypothèses associées au premier modèle (lm1) et décrire les problèmes associés.

Q8 (2 pt). On propose le code R ci-dessous. Expliquer ce qui est fait et comment peut-ont interpréter les 3 sorties numériques finales ?

```
F=matrix(0, nc=4, nr=1000)
for (i in 1:1000) {
    CO2_sim=sample(x=CO2, size=length(CO2), replace=FALSE)
     lm_temp=lm(CO2_sim~sanitaire*democratie)
    F[i,]=anova(lm_temp)$"F value"
    }
sum(F[,1]>anova(lm1)$"F value"[1])/1000
sum(F[,2]>anova(lm1)$"F value"[2])/1000
sum(F[,3]>anova(lm1)$"F value"[3])/1000
```

Q9 (2 pt). On se propose de transformer la variable CO2 en ln(CO2), refaire l'analyse avec la variable transformée et refaire les conclusions.

Q10 (1 pt). En fonction des différents modèles et des analyses réalisées jusque là, discuter de la robustesse de l'anova en fonction des conditions d'application du test.

Q11. (2 pt) Les variables consommation d'eau et émission de CO2 sont-elles corrélées ? Faire un test paramétrique et un test non paramétrique et discuter de la différence de pvalue entre les deux tests.

On s'intéresse maintenant à la géographie des émissions de CO2, avec les variables continent (considérée comme fixe) et region_monde (considérée ici, pour les besoins de l'exercice, comme aléatoire).

Q12 (1 pt). Décrire le plan d'expérience qui intègre les variables continent et region_monde pour expliquer les émissions de CO2 (la variable pays ne sera pas considérée ici).

Q13 (4 pts). Construire le modèle adéquate pour estimer la part de variabilité apportée par les pays et par les régions du monde, puis comparez globalement les différents continents. Donner ces valeurs et justifiez vos conclusions par les tests appropriés. Si les fonctions utilisées ne permettent pas d'obtenir des tests pour les effets fixes (manque de données), vous le préciserez sur votre copie et ne chercherez pas à faire d'autres analyses.