Statistiques — Modèles linéaires
M2 SIGMA — UE 905

F. Laroche

2024-2025

En vous organisant par groupe de quatre maximum, vous rédigerez un court rapport traitant les questions suivantes. Vous fournirez le code utilisé avec le rapport. Le jeu de données de travail pour le projet vous est fourni avec le sujet. Le fichier se nomme 'dataProjet_2025.csv'. Vous pouvez l'importer dans votre interface *Rstudio* via la fonction 'read.csv()'.

Question 1 — Filtrer le jeu de données pour ne garder que les charmes (valeur "Carpinus betulus L., 1753" du champ 'recherche_esp_lb_nom_plantae'; cf. support de cours). Evaluer si un modèle ANOVA sur le diamètre (DBH) en fonction de l'effet du relevé a des résidus compatibles avec les hypothèses du modèle linéaire (normalité, homoscédasticité). Si le modèle obtenu est valide, analyser et commenter les sorties du modèle (effets, intervalles de confiance, significativité, coefficient de détermination).

Question 2 — Dans le jeu de données fourni ici, on donne une nouvelle variable, la date de dernière coupe massive (champ 'lastLog'). Filtrer le jeu de données pour ne garder que les chênes (valeur "Quercus L., 1753" du champ 'recherche_esp_lb_nom_plantae'; cf. support de cours). Sur le jeu de données filtré, appliquer un modèle de régression linéaire du diamètre des charmes en fonction de cette variable 'lastLog'. Faire un diagnostic des résidus. Est-il judicieux d'introduire un terme non-linéaire de dépendance en 'lastLog' (par exemple un terme quadratique comme on a pu le faire avec "alti" dans le cours) ? Si le modèle obtenu est valide, analyser et commenter les sorties du modèle (effets, intervalles de confiance, significativité, coefficient de détermination).

Question 3 —Filtrer le jeu de données pour ne garder que les chênes (valeur "Quercus L., 1753" du champ 'recherche_esp_lb_nom_plantae'; cf. support de cours). Comme vu dans l'introduction du cours, le plan d'échantillonnage des relevés est hiérarchique dans l'espace. Ici, le jeux de données contient trois grands triangles contenant chacun trois relevés. Proposer une visualisation de cette structure. Dans le cours on a fait un modèle ANOVA pour expliquer le diamètre des chênes en utilisant le relevé comme base de définition des sous-populations. Peut-on définir plutôt des sous-populations sur la base des triplets de trois relevés, plutôt que sur chaque relevé individuellement ? Faire un modèle ANOVA correspondant à ce niveau d'agrégation, vérifier les hypothèses du modèle puis, le cas échéant, utiliser un test de modèle emboîtés pour voir si l'on peut effectivement fusionner les sous populations au sein d'un grand triangle.

Question 4 — Toujours sur le jeu de données filtré sur les chênes, dans le cours, on a mobilisé les modèles mixtes pour étudier dans quelle mesure les différences de latitude entre les relevés pouvait expliquer l'effet relevé détecté dans les modèles linéaires classiques. En reprenant la

logique de l'analyse de l'effet latitude, analyser avec un modèle mixte dans quelle mesure la variable 'lastLog' explique l'effet relevé.

Question 5 (bonus) — Toujours sur le jeu de données filtré sur les chênes, dans le cours, on a étudié avec un modèle linéaire généralisé de type binomial l'effet du relevé et de l'altitude de l'arbre sur la présence ou non d'une cavité basse. Reprenez ce modèle et ajoutez le diamètre de l'arbre en variable explicative. Vérifiez les hypothèse du nouveau modèle avec le package DHARMa en suivant l'exemple du polycopié et analysez les résultats si celles-ci sont vérifiées.

Bon courage !