Cas pratique : régression multiple sous R

Jean-Yves Barnagaud : jean-yves.barnagaud@ephe.psl.eu

Objectifs : *Construire un modèle de régression linéaire à partir d'une question biologique, l'écrire formellement sous forme d'une équation, l'implémenter sous R, le vérifier, l'interpréter et représenter les résultats.*

Cas d'étude 1: Nous allons analyser des données de mesures de Manchots adélie récoltées dans le cadre d'une étude de l'Institut Paul-Emile Victor et de l'Institut Pluridisciplinaire Hubert Curien sur la base Dumont d'Urville, en Antarctique (programme 137, mené par Céline Le Bohec). Nous allons nous intéresser à un sujet assez classique : une relation allométrique entre la longueur d'aile, le poids et le sexe des individus. Nous allons tester conjointement deux hypothèses :

Hypothèse 1: Les manchots à longues ailes sont les plus lourds. C'est la relation allométrique proprement dite, qu'on devrait observer si les variations individuelles du poids liées au contexte (âge, état de santé, état nutritif…) sont faibles relativement aux contraintes physiologiques et anatomiques ;

Hypothèse 2 : Les manchots mâles ont des ailes plus grandes que les manchots femelles. Il s'agit à la fois d'une hypothèse biologique fondée sur le dimorphisme sexuel connu chez d'autres espèces et sur une impression de terrain, et d'un possible effet confondant par rapport au résultat de la première hypothèse.

Données : Les données sont stockées dans *manchots_adelie.txt*

Les longueurs d'ailes (mm) sont des moyennes à partir de trois mesures opérées successivement par le même observateur – la variance de ces mesures n'est pas reportée, mais on dispose de l'identité de l'observateur. Les masses sont exprimées en kg.

Cas d'étude 2: Dans le cadre d'une étude sur les variations de succès reproducteur des mésanges en forêt, on s'intéresse à la quantification et à la phénologie de la ressource alimentaire principale pour les jeunes oiseaux, des chenilles phytophages qui vivent dans les houppiers des chênes et des pins. Ces chenilles produisent en permanence des fecès qui

tombent au sol : en les collectant au moyen de copromètres (des toiles d'1m² tendue sous les houppiers), on peut effectuer un suivi indirect de la ressource. Nous allons ici chercher à tester plusieurs hypothèses sur la disponibilité en chenilles.

Hypothèse 1 : La disponibilité en chenilles tend à augmenter au cours du printemps (de mars à juin)

Hypothèse 2 : La disponibilité en chenilles est plus forte en chênes qu'en pin, dans les vieux arbres, et dans les houppiers les plus volumineux (quelle que soit l'essence).

Données : Les données sont stockées dans *coprometrie_Orleans.txt*. Elles sont issues d'une thèse (Barnagaud, Univ. Orléans & IRSTEA, 2011) – nous n'en fournissons ici qu'un extrait car le jeu de données total serait trop complexe à analyser en intégralité. Concrètement, nous avons supprimé certains copromètres qui étaient redondants (plusieurs sous le même arbre, afin de tester diverses sous-questions). Vous disposez de m (la masse de fecès récoltée, proxy de la quantité de chenilles), du code du copromètre (alphanumérique), de la date (format classique et format julien en nombre de jours après le premier janvier), de l'essence d'arbre (CH : chêne ou PS : pin), du diamètre de l'arbre (cm, considéré comme proxy de son âge, même si cette approximation est sujette à caution), et du volume du houppier (m^3), calculé à partir de mesures télémétriques de hauteur et de diamètre.

Tâches à effectuer :

- Comprendre la question et la reformuler comme un modèle statistique
- Identifier les dimensions de variation des données, écrire l'équation du modèle
- Explorer les variables afin de détecter d'éventuelles difficultés
- Implémenter le modèle sous R, vérifier les conditions d'application
- Interpréter les résultats et conclure sur les hypothèses
- Préparer une restitution orale de 10 minutes maximum. Vous êtes dans un groupe de travail chargé d'analyser ces données et devez donc être clair sur vos choix méthodologiques et techniques (construction du modèle, filtrages des données) et leur adéquation aux données afin que les auditeurs puissent se faire une idée de la validité de votre travail et proposer des pistes d'amélioration. La pertinence biologique est importante à mettre en valeur, mais secondaire dans la construction de votre discours

Conseils : Appuyez-vous sur le cours et le script qui lui est associé. Commencez par traduire la question sous forme d'un modèle statistique, puis explorez les données afin de bien en connaître les caractéristiques (gamme de variabilité, distributions, corrélations). Identifiez l'existence éventuelle d'effets confondants qui justifient d'ajuster le modèle. Une fois la structure du modèle calée, implémentez-le sous R et allez droit au but (vérification et résultats). Si un problème survient, ne le négligez pas : interrogez-vous dessus, cherchez une solution si nécessaire ou développez un argumentaire qui vous autorise à ignorer le problème – des analyses complémentaires peuvent être utiles. Ne négligez pas la partie de restitution : soyez rigoureux dans votre exposé du sujet, des données, des méthodes. Expliquez clairement la structure du modèle et les résultats, en vous focalisant sur ce qui est suffisant et nécessaire

pour conclure, mais sans aller au-delà. Accordez du temps à la formulation : la précision sémantique est un gage de crédibilité.