Conclusion

Eric Marcon

17 février 2024

Synthèse

Apprentissage de R

R est un langage très versatile : il existe toujours de nombreuses façons de faire la même chose.

$\rightarrow$ Trouver son environnement, son style de codage, ses packages récurrents.

Nous avons vu :

- la syntaxe de base ;
- le tidyverse ;
- la visualisation des données ;
- l'utilisation des packages.

Nous n'avons pas vu :

- Les différents langages de R :
  - Nous avons utilisé S3, il en existe d'autres.
- L'organisation interne de R, les environnements pour comprendre les conflits de noms et la portée des variables.
- Les usages avancés de R :
  - la parallélisation pour accélérer l'exécution ;
  - l'intégration de code C++ pour encore plus de vitesse ;
  - la création de packages ;
  - la gestion du flux de travail pour mettre en cache des résultats de longs calculs.

Apprentissage de R

R (avec RStudio) est un environnement de travail en plus d'un logiciel de statistiques.

Nous avons vu :

- Comment rédiger des documents très simples (bloc-note) ou très élaborés (livre), reproductibles, indépendamment de leur format final (HTML, PDF, Word...)
- Comment utiliser git pour le contrôle de source et GitHub pour le partage, l'intégration continue et la publication.

Nous n'avons pas vu :

- D'autres types de production : site web, CV, etc ;
- Les applications R interactives avec Shiny ;
- Les outils d'enseignement.

Statistiques

Le contenu du cours est assez proche de celui de Philippe Marchand (Université du Québec en Abitibi-Témiscamingue) dont le dépôt GitHub peut servir de support rédigé.

Nous avons revu les fondamentaux de la statistique :

- Les lois de probabilité fondamentales qui permettent de faire des statistiques ;
- La loi des grands nombres qui permet de relier un échantillon à sa loi ;
- Le théorème de la limite centrale qui permet d'appliquer la loi normale à tout ce qui ne l'est pas.

Statistiques

Nous avons étudié le modèle numérique en détail :

- La régression linéaire ;
- L'Anova à un facteur ;
- Les tests classiques, dans le cadre du modèle numérique.

Nous avons vu rapidement les analyses multivariées ou méthodes d'ordination :

- L'ACP (la PCoA et l'ACM) et l'AFC ;
- Les analyses directes : RDA et CCA.

Statistiques

Nous n'avons pas vu :

- l'Anova à plusieurs facteurs :
  - triviale si les facteurs sont indépendants (équivalent à une Anova à un facteur sur les combinaisons)
  - complexe si on traite les interactions, même à deux facteurs.
- le modèle linéaire généralisé :
  - quand Y ne vaut pas β_0 en moyenne et qu'une fonction de lien est nécessaire (quand Y est entier ou compris entre 0 et 1 par exemple).
- le modèle linéaire mixte :
  - quand les observations ne sont pas indépendantes ;
  - quand un groupe de données a un effet aléatoire.
- les modèles non linéaires et imbriqués.

Nous avons utilisé l'inférence fréquentiste (par maximum de vraisemblance) mais pas l'inférence bayésienne de ces modèles.

Conclusion

Philosophie générale du cours :

- penser modèle plutôt qu'outil statistique ;
- simuler des données correspondant au modèle pour le tester avant de l'appliquer aux données réelles ;
- travailler de façon reproductible :
  - des scripts, pas de presse-bouton ;
  - le contrôle de source et l'intégration continue.
- respecter les bonnes pratiques :
  - code propre, noms de variables clairs ;
  - des paramètres plutôt que des constantes.
- documenter abondamment, réutiliser ses codes :
  - le bon code nécessite du temps mais sert longtemps.