AVANT-PROPOS

Cette seconde édition est une évolution de la version initiale publiée en 2009. Nous rappelons que cette première version s'inscrivait dans la continuation du livre *Régression : théorie et applications* paru chez Springer-Verlag (Paris). Cette nouvelle édition est plus qu'une mise à jour de la version initiale, la structure a été complètement repensée et de nouvelles parties sont apparues. Par ailleurs, un site web dédié au livre est proposé à l'url `https://regression-avec-r.github.io/`. On pourra notamment y trouver tous les jeux de données et les lignes de code utilisés dans chaque chapitre ainsi que les corrections des exercices.

L'objectif de cet ouvrage est de rendre accessible au plus grand nombre les différentes façons d'aborder un des problèmes auquel le statisticien est très souvent confronté : la *régression*. Les aspects théoriques et pratiques sont simultanément présentés. En effet, comme pour toute méthode statistique, il est nécessaire de comprendre précisément le modèle utilisé pour proposer des résultats pertinents sur des problèmes concrets. Si ces deux objectifs sont atteints, il sera alors aisé de transposer ces acquis à d'autres méthodes, moyennant un investissement modéré. Les grandes étapes – modélisation, estimation, choix de variables, examen de la validité du modèle choisi – restent les mêmes d'une méthode à l'autre. C'est dans cet esprit que cette nouvelle édition a été écrite.

Nous avons donc souhaité un livre avec toute la rigueur scientifique possible mais dont le contenu et les idées ne soient pas noyés dans les démonstrations et les lignes de calculs. Pour cela, seules quelques démonstrations, que nous pensons importantes, sont conservées dans le corps du texte. Les autres résultats sont démontrés à titre d'exercice. Des exercices, de difficultés variables, sont proposés en fin de chapitre. La présence de † indique des exercices plus difficiles. Des questions de cours sous la forme de QCM sont aussi proposées afin d'aider aux révisions du chapitre. Les corrections sont fournies sur le site du livre.

Afin que les connaissances acquises ne restent pas uniquement théoriques, nous avons intégré des exemples traités avec le logiciel libre R. Grâce aux commandes rapportées dans le livre, le lecteur pourra ainsi se familiariser avec le logiciel et retrouver les mêmes résultats que ceux donnés dans le livre. Nous encourageons donc les lecteurs à utiliser les données et les codes afin de s'approprier la théorie mais aussi la pratique.

Cet ouvrage s'adresse aux étudiants des filières scientifiques, élèves ingénieurs, chercheurs dans les domaines appliqués et plus généralement à toutes les personnes confrontés à un problème de régression. Il utilise notamment les notions de modèle, estimateur, biais-variance, intervalle de confiance, test... Pour les lecteurs peu à l'aise avec ces concepts, le livre de Lejeune (2004) pourra constituer une aide précieuse pour certains paragraphes. Cet ouvrage nécessite la connaissance des bases du calcul matriciel : définition d'une matrice, somme, produit, inverse, ainsi que valeurs propres et vecteurs propres. Des résultats classiques sont toutefois rappelés en annexes afin d'éviter de consulter trop souvent d'autres ouvrages.

Le livre se décompose en cinq parties, chacune constituée de deux à quatre chapitres. La première pose les fondamentaux du problème de régression et montre, à

travers quelques exemples, comment on peut l'aborder à l'aide d'un modèle linéaire simple d'abord, puis multiple. Les problèmes d'estimation ainsi que la géométrie associée à la méthode des moindres carrés sont proposés dans les deux premiers chapitres de cette partie. Le troisième chapitre propose les principaux diagnostics qui permettent de s'assurer de la validité du modèle tandis que le dernier présente quelques stratégies à envisager lorsque les hypothèses classiques du modèle linéaire ne sont pas vérifiées.

La seconde partie aborde la partie inférentielle. Il s'agit d'une des parties les plus techniques et calculatoires de l'ouvrage. Cette partie permet, entre autres, d'exposer précisément les procédures de tests et de construction d'intervalles de confiance dans le modèle linéaire. Elle décrit également les spécificités engendrées par l'utilisation de variables qualitatives dans ce modèle.

La troisième partie est consacrée à un problème désormais courant en régression : la réduction de la dimension. En effet, face à l'augmentation conséquente des données, nous sommes de plus en plus confrontés à des problèmes où le nombre de variables est (très) grand. Les techniques standards appliquées à ce type de données se révèlent souvent peu performantes et il est nécessaire de trouver des alternatives. Nous présentons tout d'abord les techniques classiques de choix de variables qui consistent à se donner un critère de performance et à rechercher à l'aide de procédures exhaustives ou pas à pas le sous-groupe de variables qui optimise le critère donné. Nous présentons ensuite les approches régularisées de type Ridge-Lasso qui consistent à trouver les estimateurs qui optimisent le critère des moindres carrés pénalisés par une fonction de la norme des paramètres. Le troisième chapitre propose de faire la régression non pas sur les variables initiales mais sur des combinaisons linéaires de celles-ci. Nous insistons sur la régression sur composantes principales (PCR) et la régression *Partial Least Square* (PLS). A ce stade, nous disposons de plusieurs algorithmes qui répondent à un même problème de régression. Il devient important de se donner une méthode qui permette d'en choisir un automatiquement (on ne laisse pas l'utilisateur décider, ce sont les données qui doivent choisir). Nous proposons un protocole basé sur la minimisation de risques empiriques calculés par des algorithmes de type validation croisée qui permet de choisir l'algorithme le plus approprié pour un problème donné.

Dans la quatrième partie, entièrement nouvelle, nous présentons le modèle linéaire généralisé. Cette partie généralise les modèles initiaux, qui permettaient de traiter uniquement le cas d'une variable à expliquer continue, à des variables à expliquer binaire (régression logistique) ou de comptage (régression de Poisson). Nous insistons uniquement sur les spécificités associées à ces types de variables, la plupart des concepts étudiés précédemment s'adaptent directement à ces cas nouveaux.

Enfin, la cinquième et dernière partie est dédiée à une introduction à l'estimation non paramétrique. Cette partie présente brièvement les estimateurs de type moyennes locales à travers les exemples des splines, estimateurs à noyau et des plus proches voisins. Elle inclut également une discussion sur les avantages et inconvénients d'une telle modélisation face aux modèles paramétriques étudiés précédemment.