Avant-propos

Depuis leur formulation par Nelder et Wedderburn dans le *Journal of the Royal Statistical Society : Series A* en 1972, les modèles linéaires généralisés (souvent désignés par leur acronyme anglais GLM, pour *generalized linear models*) sont devenus une des pierres angulaires de la modélisation statistique. Ils ont suscité – et suscitent encore – une littérature abondante, alimentée par des questions théoriques, méthodologiques, ou en lien avec des applications. Cet ouvrage leur est consacré. Il n'a bien sûr pas l'ambition de donner un exposé exhaustif de cette littérature. Plusieurs des thèmes abordés dans ses chapitres mériteraient en effet de faire chacun l'objet d'un ouvrage à part entière. Ainsi en est-il de la question des données manquantes, ou de la prise en compte de la censure dans les GLM. L'ambition de cet ouvrage (et de l'auteur de ces lignes) est donc plus modeste. Il s'agit ici de décrire quelques problèmes récemment étudiés dans les GLM (données manquantes, données censurées, excès de zéros) et de rapporter, encore une fois sans prétendre à l'exhaustivité, les solutions qui leur ont été apportées.

Les sujets traités dans cet ouvrage ne couvrent donc pas l'immense variété des contributions à la littérature sur les GLM. On n'y trouvera pas, par exemple, de chapitre consacré aux questions de validation de modèle, ou de sélection de variables en grande dimension. En fait, le choix des sujets abordés recouvre une part de subjectivité, et reflète en grande partie les centres d'intérêt de l'auteur. Un autre impératif a guidé la rédaction de cet ouvrage et le choix de son contenu : les méthodes qui y sont décrites peuvent, pour la plupart, être mises en œuvre à l'aide de fonctions dédiées, immédiatement disponibles dans des packages du logiciel statistique et d'analyse des données `R` (logiciel libre et gratuit). Certaines méthodes nécessitent un petit travail de programmation, et des exemples de code `R` sont donnés au fil du document.

Notons également que la plupart des problèmes décrits ici (tels que les problèmes de données manquantes ou de données censurées) ne se posent pas que dans les GLM.

Aussi avons-nous essayé de donner une description suffisamment générale des solutions proposées, afin que le lecteur en saisisse les grands principes et puisse les appliquer, ou les adapter, dans d'autres contextes. Enfin, cet ouvrage a été rédigé de telle sorte qu'il puisse être abordé à différents niveaux de lecture. Il est ainsi possible d'en comprendre le contenu sans s'attarder sur les démonstrations théoriques, pour la plupart données dans des annexes aux chapitres. Si ces chapitres sont rédigés de manière à pouvoir être lus (presque) indépendamment les uns des autres, il est toutefois recommandé de les parcourir dans l'ordre de la table des matières, qui respecte une progression croissante dans la difficulté des méthodes décrites. Un jeu de données (disponible sous R) sert de fil rouge à la rédaction de l'ouvrage, il est décrit à la section 2.5, qui devra être lue avant les autres sections utilisant ce jeu de données.

Mes remerciements les plus chaleureux vont à Nikolaos Limnios, pour ses encouragements constants depuis ma thèse, pour m'avoir invité à rédiger cet ouvrage, et suggéré des ajouts.