Introduction

À la mémoire de mon père
À ma fille Laura

Cet ouvrage est consacré aux méthodes statistiques d'analyse des données de comptage surdispersées. On parle de surdispersion lorsque la variance des observations est supérieure à celle attendue sous le modèle utilisé. Ses causes sont variées. Citons, entre autres, l'omission dans le modèle de variables explicatives importantes, la présence d'observations aberrantes ou d'une hétérogénéité inobservée entre les individus de l'échantillon.

Un excès d'observations égales à zéro (on parle d'« excès de zéros » ou d'« inflation de zéros ») est une autre cause majeure de la surdispersion, que l'on rencontre depuis les domaines d'application les plus « classiques » : assurance (automobile en particulier, en raison du système de bonus-malus), économie (de la santé par exemple, dans l'étude de la consommation de soins et des déterminants du renoncement aux soins médicaux), épidémiologie, sociologie (étude des déterminants du non-recours aux prestations sociales), etc. jusqu'aux plus inattendus, tels que l'étude des conflits et du terrorisme.

Si les premiers modèles statistiques pour données de comptage en présence d'inflation de zéros ont été développés dès les années 1960, ce n'est que depuis le début des années 2000 que des modèles réellement élaborés ont commencé à être proposés. Citons toutefois trois jalons, précurseurs, posés par Mullahy [MUL 86], Lambert [LAM 92] et Hall [HAL 00]. Ces auteurs ont proposé les premiers modèles *de régression* à inflation de zéros et ouvert la voie à de nombreux travaux méthodologiques, théoriques et appliqués.

Rappelons que l'objet de la régression est de modéliser et étudier la relation entre une variable, dite variable réponse, et une ou plusieurs variables dites explicatives.

Le modèle de régression linéaire constitue le point d'entrée habituel dans le monde de la régression, suivi généralement par les modèles de comptage classiques (modèle de régression binomial et de Poisson, pour ne citer que les plus connus), souvent vus dans le cadre unifiant des modèles linéaires généralisés, introduits par Nelder et Wedderburn [NEL 72]. L'enseignement de ces modèles est devenu incontournable dans les formations d'ingénierie statistique, d'économétrie, de biostatistique, etc. et plus généralement dans les formations universitaires et d'ingénieurs qui intègrent de la modélisation statistique. Les méthodes de prise en compte de la surdispersion (méthodes de correction de la variance, modèles à inflation de zéros) sont en revanche moins étudiées.

L'objectif de cet ouvrage est de fournir une introduction à ces méthodes, qu'il ne couvre évidemment pas de manière exhaustive : le sujet est trop vaste ! Son ambition, beaucoup plus modeste, est de fournir au lecteur quelques outils, parmi les plus classiques, de prise en compte de la surdispersion dans les données de comptage. Ces outils sont présentés de manière rigoureuse, mais sans formalisme excessif et l'ouvrage est accessible à tout lecteur, statisticien de formation ou non, qui maîtrise les notions usuelles d'algèbre linéaire et de probabilités nécessaires à l'enseignement de la modélisation statistique. La lecture des sections consacrées à l'asymptotique dans les modèles de régression à inflation de zéros peut être omise sans gêner la compréhension générale de l'ouvrage.

Pour rendre plus concrets les outils et méthodes décrits dans cet ouvrage, nous les avons accompagnés d'applications sur des données réelles, traitées à l'aide du logiciel statistique libre et gratuit `R` [RCO 17]. Un même jeu de données sert de fil rouge tout au long des chapitres 2 à 4. Ces données, issues d'une enquête nationale sur les dépenses médicales réalisée aux USA, sont disponibles dans le package `AER` [KLE 08] de `R`. Notre objectif est d'illustrer les méthodes et modèles statistiques permettant de prendre en compte la surdispersion, aussi n'avons-nous pas cherché à construire le modèle explicatif le plus fin possible, en recherchant par exemple des interactions entre variables explicatives. Nous ne nous sommes pas non plus intéressés à la question de la validation des modèles présentés et laissons le lecteur intéressé par cet aspect (utilisation des résidus, mesures d'influence, outils diagnostics, etc.) se tourner vers les ouvrages (ils sont nombreux !) beaucoup plus compétents sur ces sujets.

Les modèles de régression à inflation de zéros constituent une partie importante de l'ouvrage. Le chapitre 4 leur est entièrement consacré. La maîtrise d'un minimum de notions sur les modèles linéaires généralisés (formalisme des modèles, construction d'estimateurs et de tests, aspects numériques) est requise avant d'en aborder la lecture. Et pour s'imprégner au mieux de ces notions, il nous a semblé judicieux de revenir aux concepts de base et aux résultats essentiels de la régression linéaire. Le plan de l'ouvrage est donc le suivant. Le chapitre 1 est consacré au modèle de régression linéaire. Le chapitre 2 présente les modèles linéaires généralisés et met l'accent sur les modèles de régression binomial et de Poisson. Le chapitre 3 introduit la question de la

surdispersion dans les données de comptage et présente quelques outils et modèles statistiques permettant de la prendre en compte. Le chapitre 4 traite plus particulièrement des modèles de régression à inflation de zéros. Une partie relativement importante de ce chapitre est consacrée à des avancées récentes dans le domaine. En particulier, nous nous intéressons à la question de l'identifiabilité dans ce type de modèles et à ses implications pratiques en termes de construction de modèle. Nous nous intéressons également aux propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans ces modèles. Si cette partie est plus théorique et sans doute plus aride à lire, elle ne doit pas apparaître moins importante aux yeux du praticien de la statistique, car elle justifie le recours aux méthodes usuelles d'inférence statistique basées sur des arguments asymptotiques, telles qu'elles sont mises en œuvre, par exemple, dans les modèles linéaires généralisés.