Modélisation et d'analyse dans les études statistiques

Bernard CLÉMENT, PhD

L'idée fondamentale de l'analyse statistique est basée sur la modélisation mathématique entre des variables, l'identification de leurs rôles et l'objectif visé par l'étude. On distingue entre les **études exploratoire** d'ensembles de données historiques ou observationnelles et les **études confirmatoires** dont les données provenant d'expériences planifiées. On fait aussi la distinction entre des **études de nature énumérative** dont les données provienent d'un échantillonnage aléatoire d'une population fixe d'unités statistiques comme dans les sondages et les enquêtes. Il y a aussi les **études analytiques** dont les données sont le résultat d'observations sur un processus possiblement dynamique dans le temps. Dans ce dernier cas le concept de population n'est pas parfaitement défini.

La stratégie de **modélisation statistique** d'une variable de réponse Y en fonction de plusieurs variables (facteurs explicatifs) X1, X2,… est une **approche globale du haut vers le bas (« top down »)**. On commence avec un modèle général complet qui tient compte des effets principaux et des effets d'interaction de toutes les variables explicatives. On distingue **2 catégories de modèles statistiques** : les **modèles de régression et les modèles d'analyse de la variance.**

Dans les **modèles de régression** on cherche à établir une équation pour faire de la prédiction, de l'interpolation, de l'extrapolation ou pour faire de l'optimisation de la réponse (maximisation, minimisation, cible). On recherche un modèle ou quelques modèles parcimonieux avec un fort pouvoir explicatif. Les méthodes de sélection de variables sont employées pour identifier les effets significatifs et importants. On dégage quelques modèles, les plus simples possibles, en éliminant les effets principaux, et d'interaction qui ne jouent aucun rôle pour expliquer la réponse. On recherche un ou des **modèles parcimonieux** ayant un nombre relativement petit d'effets significatifs et importants sur la réponse.

Dans certaines d'études, on cherche simplement à établir l'influence réelle (significative / importante) ou non des facteurs et leurs effets sur la réponse. Dans ce dernier cas, le modèle développé n'a pas besoin d'avoir un fort pouvoir explicatif car l'équation de prédiction n'est pas exploitée. C'est le domaine des **modèles d'analyse de variance et de covariance.** Le modèle sert à montrer si certains effets (principaux, interaction) ont un signal relativement fort en comparaison de l'erreur expérimentale. Le signal est mesuré avec un ratio ou rapport-signal bruit. Le bruit, aussi appelé **erreur expérimentale**, représente l'effet combiné de tous les facteurs connus ou inconnus qui ne sont pas explicitement tenu en compte dans les effets des variables explicatives identifiées.

L'erreur expérimentale est une source de variabilité et une composante additionnelle toujours présente dans tous les **modèles statistiques**. On reconnait explicitement la présence de l'erreur expérimentale dans l'observation et la collecte de données (échantillonnage), dans le modèle et son développement. Cette caractéristique fondamentale est prise en compte dans le processus de modélisation. C'est ce qui distingue les modèles statistiques de toutes les autres méthodes de modélisation mathématique pour décrire les phénomènes observés.

Les **étapes** du processus de **modélisation statistique** sont :

- **Identification** **processus/problème**
- **Observation** **plan collecte des données**
- **Spécification** **modèle pour analyse**
- **Estimation** **paramètres du modèle**
- **Décomposition** **variabilité**
- **Validation** **analyse résiduelle**
- **Exploitation** **optimisation / résolution problème**