Processus de modélisation statistique

Bernard CLÉMENT, PhD

La **stratégie de choix** recommandée pour faire la **modélisation statistique** d'une variable de réponse Y en fonction de plusieurs facteurs explicatifs X1, X2,… est une **approche globale du haut vers le bas.** On commence avec un modèle général complet qui tient compte des effets principaux et des effets d'interaction de toutes les variables explicatives.

Après une suite d'analyses et de méthodes de sélection de variables et des effets qui s'avèrent significatifs et importants, on dégage quelques modèles, les plus simples possible, en éliminant les effets (principaux, interaction) qui ne jouent aucun rôle pour expliquer la réponse. On recherche un ou des **modèles parcimonieu**x ayant un nombre relativement petit d'effets significatifs et importants sur la réponse. Si on veut exploiter le modèle pour faire de l'interpolation ou de l'extrapolation ou pour faire de l'optimisation de la réponse (maximisation, minimisation, cible) on recherche un modèle ayant un fort pouvoir explicatif. C'est le domaine des **modèles de régression.**

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 mathématiques pour décrire les phénomènes observés.

Le **processus de modélisation statistique** est composé des étapes suivantes :

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