Introduction au Modèle Linéaire

Résumé

Introduction au modèle linéaire ou de régression et au modèle linéaire général.

Plan du cours :

- *Régression linéaire simple*
- *Modèle gaussien : régression linéaire multiple*
- *Modèle gaussien : analyses de variance et covariance*
- *Modèle binomial ou régression logistique*
- *Modèle poissonnien ou loglinéaire*
- *Introduction au modèle linéaire général*

Objectifs

La *Statistique* a plusieurs objets : descriptif ou exploratoire, décisionnel (tests), modélisation selon que l'on cherche à représenter des structures des données, confirmer ou expliciter un modèle théorique ou encore prévoir. Ce cours s'intéresse au thème de la *modélisation* et plus particulièrement aux méthodes *linéaires* et à celles qui se ramènent au cas linéaire. Il se limite donc à l'exposé des méthodes dites *paramétriques* dans lesquelles interviennent des *combinaisons linéaires* des variables dites explicatives. Celles-ci visent donc à l'estimation d'un nombre généralement restreint de paramètres intervenant dans cette combinaison mais sans aborder les techniques spécifiques à l'étude des séries chronologiques. Les méthodes non-paramétriques élémentaires (loess, noyaux, splines) seront introduites dans le cas unidimensionnel.

Le *cadre général* de ce cours considère donc les observations d'une variable aléatoire Y dite *réponse, exogène, dépendante* qui doit être expliquée (modélisée) par les mesures effectuées sur p variables dites *explicatives, de contrôle, endogènes, dépendantes, régresseurs*. Ces variables peuvent être quantitatives ou qualitatives, ce critère déterminant le type de méthode ou de modèle à mettre en œuvre : régression linéaire, analyse de variance et covariance, régression logistique, modèle log-linéaire.

Compte tenu du temps limité et de la variété des outils mis en jeu nous avons fait le choix d'insister sur la *pratique* des méthodes considérées ainsi que sur la compréhension des sorties proposées par un logiciel (SAS/STAT) et de leurs limites plutôt que sur les fondements théoriques. Ce cours s'inspire largement d'une présentation "anglo-saxonne" de la Statistique, du particulier vers le général, dont des compléments sont à rechercher dans la bibliographie citée en référence. On montre donc comment utiliser les propriétés des modèles statistiques pour le traitement des données tandis que certains des aspects plus mathématiques (démonstrations) sont l'objet d'exercices. Néanmoins, le dernier chapitre introduit au cadre théorique général incluant toutes les méthodes considérées : le *modèle linéaire généralisé*.

En théorie, on peut distinguer deux approches : avec ou sans hypothèse probabiliste sur la distribution des observations ou des erreurs qui est, le plus souvent, l'hypothèse de *normalité*. En pratique, cette hypothèse n'est guère prouvable, les tests effectués sur les résidus estimés sont peu puissants. Cette hypothèse est néanmoins implicitement utilisée par les logiciels qui produisent systématiquement les résultats de tests. Plus rigoureusement, ces résultats sont justifiés par les propriétés des distributions asymptotiques des estimateurs, propriétés qui ne sont pas développées dans ce cours. En conséquence, du moment que les échantillons sont de taille "raisonnable", hypothèse on non de normalité, les distributions des estimateurs et donc les statistiques de test sont considérées comme valides.

En revanche, d'autres aspects des hypothèses, inhérentes aux méthodes développées et qui, en pratique, conditionnent fortement la qualité des estimations, doivent être évalués avec soin : *linéarité, colinéarité, homoscédasticité, points influents* ou atypiques (outliers). Les différents *diagnostics* ainsi que le problème du choix des variables explicatives, c'est-à-dire du *choix de modèle*, sont plus particulièrement décrits.

Dans la mesure du possible, nous avons respecté une certaine uniformisation des notations. Des caractères majuscules X, Y désignent des variables aléatoires, des caractères gras minuscules désignent des vecteurs : y_i est la $i$ème observation de Y rangée dans le vecteur $\mathbf{y}$, un chapeau désigne un prédicteur : $\hat{y}_i$, les caractères gras majuscules sont des matrices, un caractère grec (β) est un paramètre (qui est une variable aléatoire) dont l'estimation est désignée par la lettre latine correspondante (b).