Régression linéaire symbolique avec variables taxonomiques.

Filipe Afonso*·**, Lynne Billard***
Edwin Diday*

*Ceremade/Université Paris 9 Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16, France.
afonso@ceremade.dauphine.fr
diday@ceremade.dauphine.fr
**Lamsade/ Université Paris 9 Dauphine
***Department of statistics/University of Georgia
Athens, 30602, USA.
lynne@stat.uga.edu

Résumé. Le présent papier concerne l'extension des méthodes classiques de régression linéaire aux cas des données symboliques et fait suite à de précédents travaux de Billard et Diday sur la régression linéaire avec variables intervalles et histogrammes. Dans ce papier, nous présentons des méthodes de régression avec variables taxonomiques. Les variables taxonomiques sont des variables organisées en arbre exprimant plusieurs niveaux de généralité (les villes sont regroupées en régions qui sont elles-mêmes regroupées en pays). La méthode proposée sera testée sur données simulées. Finalement, nous observerons que ces méthodes nous permettent d'utiliser la régression linéaire pour étudier des concepts et pour réduire le nombre de données afin d'améliorer les résultats obtenus par rapport à une régression classique.

1 Introduction

Dans la pratique, nous sommes souvent intéressés par l'étude de groupes d'individus plutôt que les individus statistiques eux-mêmes. Aussi, les bases de données atteignent des masses considérables d'observations et la réduction du nombre de données peut faciliter les études. Dans ces deux cas une agrégation des données va nous amener à manipuler des variables qui ne sont pas à valeurs uniques. Nous obtenons par exemple des intervalles, des histogrammes et des diagrammes. De plus, ces variables peuvent être organisées par des taxonomies ou des hiérarchies. Les données ainsi constituées sont appelées données symboliques (Billard et Diday 2003, Bock et Diday 2000). Des travaux sur l'extension des méthodes de régression linéaire aux cas des variables intervalles et histogrammes ont déjà été entrepris dans (Billard et Diday 2000 et 2002). Dans ce papier, nous nous intéressons aux variables taxonomiques (Voir aussi Afonso et al 2003).

2 Problématique

En régression linéaire, nous voulons expliquer une variable dépendante Y à partir de k variables explicatives $X_1,\ldots, X_k$ sous la forme d'un modèle linéaire $Y= a+b_1X_1+\ldots+b_kX_k+\varepsilon = \beta X+\varepsilon$ où ε constitue le résidu. Dans la théorie classique, nous calculons le vecteur optimal β^*