ELSEVIER

Computational Statistics & Data Analysis 34 (2000) 371–386

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Combining non-parametric models with logistic regression: an application to motor vehicle injury data

Petra M. Kuhnert[a,*], Kim-Anh Do[b], Rod McClure[c]

[a] *School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia*

[b] *Department of Biostatistics, Box 213, University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030,USA*

[c] *CONROD, Department of Medicine, Royal Brisbane Hospital, Herston, QLD 4029, Australia*

Received 1 September 1999; received in revised form 1 November 1999

Abstract

To date, computer-intensive non-parametric modelling procedures such as classification and regression trees (CART) and multivariate adaptive regression splines (MARS) have rarely been used in the analysis of epidemiological studies. Most published studies focus on techniques such as logistic regression to summarise their results simply in the form of odds ratios. However flexible, non-parametric techniques such as CART and MARS can provide more informative and attractive models whose individual components can be displayed graphically. An application of these sophisticated techniques in the analysis of an epidemiological case-control study of injuries resulting from motor vehicle accidents has been encouraging. They have not only identified potential areas of risk largely governed by age and number of years driving experience but can also identify outlier groups and can be used as a precursor to a more detailed logistic regression analysis.

Keywords: Classification and regression trees; Injury; Logistic regression; Multivariate adaptive regression splines; Recursive partitioning

* Corresponding author. Tel.: +61-7-3864-5267; fax: +61-7-3864-2310.
E-mail address: p.kuhnert@fsc.qut.edu.au (P.M. Kuhnert).

0167-9473/00/$ - see front matter
PII: S 0167-9473(99)00099-7

1. Introduction

1.1. Background

A common problem of most practical research is to assess relationships among a set of variables. A primary statistical tool is regression analysis which may be used to evaluate the relationship of one or more covariates or predictor variables $x_1,\ldots,x_n$ to a single (continuous or binary/ordinal) response variable y. It is most often used when the predictor variable cannot be controlled as when collected in a sample survey or other observational study. However, regression analysis may also be applied to controlled experimental situations. The most common situations where regression analysis is appropriate include problems where one wishes to capture the joint predictive relationship of y on a small subset of $x_1,\ldots,x_n$ in the form

$$y = f(x_1,\ldots,x_k) + \varepsilon \quad (k \leq n),$$

where ε is an additive stochastic component with zero expectation. Existing methods of regression analysis range from the simple linear and polynomial regression to logistic regression (Cox, 1970; Hosmer and Lemeshow, 1989), generalised additive modelling (Hastie and Tibshirani, 1990) and recently include methods such as classification and regression trees (CART) (Breiman et al., 1984) and a sophisticated, flexible regression technique, multivariate adaptive regression splines (MARS) (Friedman, 1991) based on recursive partitioning strategies. Examples of MARS models being used in practice are limited. This is mainly due to the computational requirements of fitting a MARS model and the complexity of the resulting fit. An area which yields some applications using MARS is in medicine (Friedman and Roosen, 1995; Gill et al., 1996). Other fields taking some interest in the MARS methodology is data mining (Stone et al., 1997) where problems of dealing with large datasets arise and spatial problems such as oceanography, where modelling sea ice distribution in the southern ocean was of primary interest (De Veaux et al., 1993a). Other papers have focussed on comparisons between MARS and other non-linear regression methods (Frank, 1995; Marshall et al., 1994) and neural networks (De Veaux et al., 1993b; Ripley, 1994; Cas and Stone, 1996). MARS has even been applied to some time series applications providing improvements over existing techniques (Lewis and Stevens, 1991). A comprehensive tutorial on applying MARS to a calibration problem in chemistry is also worth reading for a gentle introduction to the technique (Sekulic and Kowalski, 1992).

1.2. Modelling in epidemiology

MARS models are rarely used in epidemiological studies, despite valuable contributions they may provide. Techniques such as CART are becoming more widely used due to their ease of interpretation and ability of handling missing data but sometimes lack the flexibility required to model all aspects of the data. The majority of applications using CART have arisen from the field of medicine where decision making is a large component of the clinician's work practice (Gill et al., 1996;