Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Computational Statistics & Data Analysis 50 (2006) 3464–3485
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs

W. Sauerbrei[a,*], C. Meier-Hirmer[b], A. Benner[c], P. Royston[d]

[a] *Institute of Medical Biometry and Medical Informatics, University Hospital of Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany*
[b] *SNCF, Paris, France*
[c] *Deutsches Krebsforschungszentrum, Heidelberg, Germany*
[d] *MRC Clinical Trials Unit, London, UK*

Received 3 September 2004; received in revised form 13 July 2005; accepted 21 July 2005
Available online 10 August 2005

Abstract

In fitting regression models data analysts are often faced with many predictor variables which may influence the outcome. Several strategies for selection of variables to identify a subset of 'important' predictors are available for many years. A further issue to model building is how to deal with non-linearity in the relationship between outcome and a continuous predictor. Traditionally, for such predictors either a linear functional relationship or a step function after grouping is assumed. However, the assumption of linearity may be incorrect, leading to a misspecified final model. For multivariable model building a systematic approach to investigate possible non-linear functional relationships based on fractional polynomials and the combination with backward elimination was proposed recently. So far a program was only available in Stata, certainly preventing a more general application of this useful procedure. The approach will be introduced, advantages will be shown in two examples, a new approach to present FP functions will be illustrated and a macro in SAS will be shortly introduced. Differences to Stata and R programs are noted.

Keywords: Multivariable model building; Function selection; Fractional polynomials; Programs

* Corresponding author. Tel.: +49 761 203 6669; fax: +49 761 203 6677.
E-mail address: wfs@imbi.uni-freiburg.de (W. Sauerbrei).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.07.015

1. Introduction

In fitting regression models data analysts are often faced with many predictor variables which may influence the outcome. Strategies for selection of variables are used to identify a subset of 'important' predictors. Difficulties associated with strategies such as sequential procedures (e.g. stepwise or backward procedures) or all-subset selection with different optimization criteria (e.g. Akaike (AIC) or Bayesian (BIC) information criteria) are overfitting, underfitting, biased estimates of the regression parameters of the final model and a lack of reproducibility of the regression parameters in new data (Miller, 1990); some approaches to investigate these issues by resampling methods are discussed in Sauerbrei (1999). Although subject-matter knowledge should guide selection, some variables will inevitably be chosen mainly by statistical principles—typically by P-values for including or excluding variables. The definition of a 'best' strategy to produce a model which has good predictive properties in new data is difficult. A model which fits the current data set well may be too much data driven to give adequate predictive accuracy in other settings.

A second obstacle to model building is how to deal with non-linearity in the relationship between outcome and a continuous or ordered predictor. The traditional assumption of linearity may be incorrect, leading to a misspecified final model in which a relevant variable may not be included because its true relationship with outcome is non-monotonic, or in which the assumed functional form differs substantially from the unknown true form. Alternatively, continuous predictors may be converted into categorical variables by grouping into two or more categories. With dichotomization, considerable variability may be subsumed within each group. The implicit model is unrealistic, since individuals close to but on opposite sides of the cutpoint have very similar rather than very different outcomes. The arbitrariness of the choice of cutpoint may encourage a search for a value which gives the most 'satisfactory' result. Taken to extremes, all possible cutpoints may be tried and the value which maximizes statistical significance may be chosen. Because of multiple testing, the overall Type I error rate will be around 40% rather than the nominal 5% (Altman et al., 1994; Miller and Siegmund, 1982; Lausen and Schumacher, 1996). The cutpoint chosen will have a wide confidence interval and will have no substantive meaning. Crucially, the difference in outcome between the two groups will be overestimated and the confidence interval will be too narrow.

An alternative approach is to keep the variable continuous and to allow some form of non-linearity. Instead of using quadratic or cubic polynomials, a general family of parametric models have been proposed by Royston and Altman (1994), that is based on so-called fractional polynomial (FP) functions. Here, usually one or two terms of the form X^p are fitted, the exponents p being chosen from a small predefined set S of integer and non-integer values. Although only a small number of functions is considered (besides no transformation ($p = 1$)), the set S includes 7 transformations for FPs of degree 1 (FP1) and 36 for FPs of degree 2 (FP2), FP functions provide a rich class of possible functional forms leading to a satisfactory fit to the data in many situations. Royston and Altman (1994) dealt mainly with the case of a single predictor, but they also suggested and illustrated an algorithm for fitting FPs in multivariable models. By combining backward elimination (BE) with the search for the most suitable FP transformation for continuous predictors Sauerbrei and Royston (1999) propose modifications to this multivariable FP (MFP) procedure. A further