X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

NPRegFast: Analyzing factor by curve interactions

Nora M. Villanueva, Marta Sestelo and Javier Roca
Department of Statistics and O. R, University of Vigo
nmvillanueva@uvigo.es, sestelo@uvigo.es, roca@uvigo.es

ABSTRACT

In the nonparametric regression framework, issues of interest include the so-called factor-by-curve interaction, where the effect of a continuous covariate on response varies across groups defined by levels of a categorical variable. This study introduces a software application for R which performs inference in a nonparametric regression model. It describes the capabilities of the program for estimating these models (and their derivatives), for drawing different regression curves by factor levels and for drawing inferences about critical points (for instance, maxima or change points). Bootstrap methods were implemented to draw inferences from the derivative curves, and binning techniques were applied to speed up computation in the estimation and testing processes. The software is illustrated using biological data.

Keywords: nonparametric regression, factor-by-curve interaction, R-project

1. INTRODUCTION

In many biological studies, the relationship between the response variable Y and the covariate X can be estimated using local polynomial estimators (Wand and Jones, 1995), resulting in a model of type

$$Y = m(X) + \varepsilon \tag{1}$$

where m is a smooth and unknown function and ε is the error which is assumed to have mean zero.

A generalization of the previous model in (1) is a model with factor-by-curve interactions, where the relationship between Y and X can vary among subsets defined by levels of a categorical covariate Z. Thus, the estimated curve will take different functional forms depending on the level j, for $j = 1, \ldots, M$ of the factor Z.

This paper releases the possible uses of the package with an optimized methodology adapted to various disciplines that wish to estimate the relationship between two continuous variables, influenced or not by a certain factor.

2. PACKAGE DESCRIPTION

The software provides a set of functions for estimating different nonparametric regression models, first and second derivatives, maximum points, etc. `NPRegFast` has included the calculation of bootstrap confidence intervals (Efron, 1979; Efron and Tibshirani, 1993). These confidence intervals will allow us to make inference from original data. This software is intended to be used with the R statistical program (R Development Core Team, 2009).

Our package is composed of several functions that enable users to fit the proposed models. Table 1 provides a summary of some objects in this package.

Table 1: Summary of functions in the package.

Function	**Description**
`frfast`	Main function for fitting regression models and obtaining the different outputs (model estimates, first and second derivatives).
`summary.frfast`	Summary of the model for the objects of class `frfast`.
`plot.frfast`	A function that provides the plots for model estimates and their confidence intervals based on bootstrap techniques.
`points.frfast`	Provides a table with the covariate X values that maximize the initial estimation, first and second derivatives.
`diff.frfast`	A function that draws the plots for differential model curves and their confidence intervals based on bootstrap methods.
`contrast.frfast`	A function that shows a table with test's results between curves.

Table 2 provides a summary of the arguments in the `frfast` function.

Table 2: Summary of main arguments of the `frfast` function.

`frfast` **arguments**	**Description**
`f`	Factor. If is `NULL`, main function `frfast` estimates the model without considering factors.
`x`	Covariable X.
`y`	Response variable Y.
`w`	Weights.
`h`	Bandwith.
`p`	Degree polynomial.
`kbin`	Number of nodes.
`nboot`	Number of bootstrap repeats.

3. APPLICATION EXAMPLE

Our methodology was used to determinate the ideal size of capture of stalked barnacle, *Pollicipes pollicipes* (Gmelin, 1789) on the Atlantic coast of Galicia (NW Spain). We have a sample size of 16600 individuals, and measures of two biometric variables of each specimen: rostro-carinal length (RC) and dry weight (DW), obtained over two years (2006-2007).

To ascertain whether the size remained constant across time and was not altered by any possible annual variability in the growth of this species, the study was broken down by year (Figure 1). In both cases, the initial regression curves show the way in which smaller-sized individuals increased in weight exponentially whereas larger-sized individuals increased in weight proportionally. The first derivatives of these curves increased as individuals grew in size, until they peaked at an RC of 21.18 mm (19.75, 23.56) in 2006 and 21.10 mm (19.60, 22.89) in 2007 (Figure 1).

To asses whether the regression curves of the two years of the study are the same, we have constructed the differential curve showed in the Figure 2 (first column). The same procedure was carried out with the first derivatives (Figure 2, second column). In both cases, we can clearly observe how the interval confidences of the two curves, in few occasions, include zero. It suggests both regression curves and their first derivatives are different between each other.

4. CONCLUSION

This paper discusses implementation in R some newly developed methods in non-parametric regression models. We note that the software returns an object of class `frfast` with components that allow users to perform a detailed analysis of these models.

REFERENCES

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26.

Efron, E. and Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Chapman and Hall, London.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

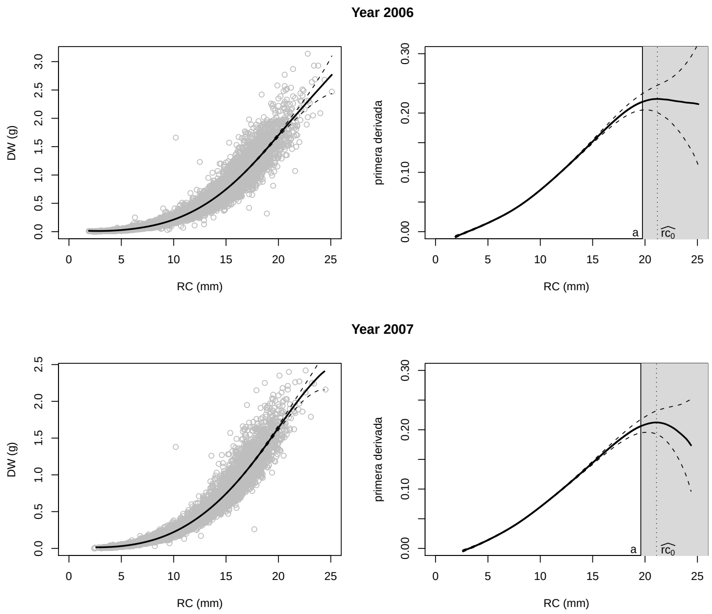

Figure 1: Regression curves and first derivatives (solid lines) with bootstrap-based 95% confidence intervals (broken lines) for dry weight and rostro-carinal length. First row: year 2006; second row: year 2007.

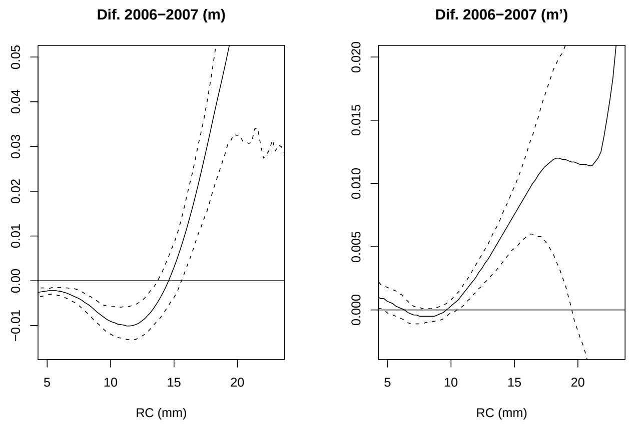

Figure 2: First column: differences (solid lines) between initial regression curves (m) of 2006-2007, with bootstrap-based 95% confidence intervals (broken lines). Second column: differences (solid lines) between first derivatives (m^1) of 2006-2007, with bootstrap-based 95% confidence intervals (broken lines).