Journal of Statistical Software

December 2017, Volume 82, Issue 12. *doi: 10.18637/jss.v082.i12*

npregfast: An R Package for Nonparametric Estimation and Inference in Life Sciences

Marta Sestelo
University of Minho

Nora M. Villanueva
University of Vigo

Luis Meira-Machado
University of Minho

Javier Roca-Pardiñas
University of Vigo

Abstract

We present the R **npregfast** package via some applications involved with the study of living organisms. The package implements nonparametric estimation procedures in regression models with or without factor-by-curve interactions. The main feature of the package is its ability to perform inference regarding these models. Namely, the implementation of different procedures to test features of the estimated regression curves: on the one hand, the comparisons between curves which may vary across groups defined by levels of a categorical variable or factor; on the other hand, the comparisons of some critical points of the curve (e.g., maxima, minima or inflection points), studying for this purpose the derivatives of the curve.

Keywords: regression, nonparametric, kernel, factor-by-curve interaction, testing, R.

1. Introduction

Regression analysis plays a fundamental role in statistics. The purpose of this technique is to evaluate the influence of some explanatory variables on the mean of the response. In the case of nonparametric regression, the dependence between the response and the covariates is modeled without specifying in advance the function that links them. Development and implementation of different methods for estimation and inference regarding these models is the central focus of this work.

Nonparametric methods are now widely recognized as useful tools in regression analysis. However, they are much more computationally demanding than their parametric counterparts. In view of the high cost entailed we used Fortran (Gehrke 1995) as the programming language. To facilitate the use in practice of the methodologies proposed, a user-friendly R (R Core Team

2017) package is implemented containing the Fortran code.

A range of methods can be found in the **npregfast** package (Sestelo, Villanueva, and Roca-Pardiñas 2017) including estimation of the conditional mean and the derivatives with/without factor-by-curve interactions, bandwidth selection and computational acceleration. In addition, several procedures to test different features of the estimated regression curves have also been included. These developments have been applied to a couple of real data situations in a life science context.

The effect of a continuous covariate on the response may vary across groups defined by levels of a categorical variable. This means that the continuous covariate can behave in a different way in the absence/presence of a factor, producing the corresponding factor-by-curve effect. Following this, there exist many practical situations that call for comparisons of regression curves and their derivatives which may vary across groups defined by different experimental conditions. Thus, the interest might be focused on drawing inferences about some critical points of the curve (e.g., maxima, minima or inflection points), studying for this purpose the derivatives of the curve. In marine biology, for example, the growth of commercial species collected in different zones could be analyzed and compared with each other taking into account the environmental conditions prevailing in such areas. The location would be considered as the factor, and the length-weight relationship could be studied including the factor-by-curve interaction. Similarly, the first derivative of the regression curve could be calculated, thereby enabling the different stages of growth to be defined as the species increases in size. Furthermore, calculation of this derivative could have a direct application in the management of this species, possibly in estimating a size of capture (Sestelo and Roca-Pardiñas 2011; Bidegain, Sestelo, Roca-Pardiñas, and Juanes 2013; Bidegain, Guinda, Sestelo, Roca-Pardiñas, Puente, and Juanes 2015).

This paper describes the R based **npregfast** package which is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=npregfast/`). The package allows to estimate the regression curves and their derivatives, to compare them between levels (in the case of including a categorical variable), and even to compare their critical points. The main estimation procedure is based on local polynomial kernel smoothers (Wand and Jones 1995; Fan and Gijbels 1996). It is also possible, however, to estimate the models using a classical parametric model – the allometric model – one of the most frequently used models in fishery management. In addition, the package implements the following two tests: (i) a global test for the equality of M regression curves; and, (ii) a local test to draw inferences about critical points linked to the derivative curves. Inference with this package (confidence intervals and tests) is based on bootstrap resampling methods (Efron 1979; Wu 1986; Liu 1988; Efron and Tibshirani 1993; Härdle and Mammen 1993; Mammen 1993; Kauermann and Opsomer 2003). Accordingly, binning acceleration techniques are also implemented to ensure that the package is computationally efficient (Fan and Marron 1994).

There are a number of contributed packages available for R, for example on CRAN, which are devoted to nonparametric estimation procedures. Particularly, a brief review of software developments for carrying out kernel-based regression could start with the `ksmooth` function of the **stats** package, which allows to obtain estimates using the Nadaraya-Watson estimator. However, to study kernel-based nonparametric estimators in depth, the **KernSmooth** package (Wand 2015) affords more possibilities for R users. Using its main function, `locpoly`, a probability density function, a regression function or their derivatives can be estimated using local polynomials. Another option might be the use of the **kerdiest** package (del Río

and Estévez-Pérez 2012), which has been designed for computing kernel estimators of the distribution function, or the **lokern** package (Herrmann and Maechler 2016), which features kernel regression smoothing with adaptive local or global plug-in bandwidth selection. In a specific context, such as item response theory or graduating mortality rates, these kernel smoothers could be applied by means of the packages **KernSmoothIRT** (Mazza, Punzo, and McGuire 2014) or **DBKGrad** (Mazza and Punzo 2014), respectively. Finally, in a multivariate framework, the **np** package (Hayfield and Racine 2008; Racine and Hayfield 2017) provides a variety of nonparametric (and semiparametric) kernel methods that seamlessly handle a mix of continuous, unordered and ordered factor data types often encountered in applied settings.

With respect to testing procedures, it is worth noting that a vast literature exists about the comparison of regression functions. Relevant papers about this topic are Hall and Hart (1990); Härdle and Marron (1990); Delgado (1993); Kulasekera (1995); Young and Bowman (1995); Bowman and Young (1996); Dette and Neumeyer (2001); Neumeyer and Dette (2003); Pardo-Fernández, Van Keilegom, and González-Manteiga (2007); Park and Kang (2008); Srihera and Stute (2010), among others. For a detailed review see González-Manteiga and Crujeiras (2013). In addition, when the previous division into groups is governed by a discrete variable, tests for the significance of this discrete variable could also be considered. Related work includes Racine, Hart, and Li (2006); Lavergne (2001) and the references therein, for example. All the above references focus on the regression functions, however, to our knowledge, there are no references dealing with the comparison of derivatives. Furthermore, there exist procedures described in the literature that test the monotonicity of the regression function (e.g., Bowman, Jones, and Gijbels 1998) or techniques as the SiZer method (Chaudhuri and Marron 1997) which let us evaluate if the observed features are really significant. This latter technique was implemented in the **SiZer** package (Sonderegger 2012). However, to the best of our knowledge, this is the first contribution dealing with the topic of testing critical points between curves.

In this article we explain and illustrate how numerical and graphical output for all methods can be obtained using the **npregfast** package via life science applications. The applications are chosen to solve two real problems related to the management of an aquatic living resource, and to the spurt in growth for school-aged children and adolescents.

The remainder of the paper is structured as follows: Section 2 describes the estimation procedures, jointly with practical questions such as bandwidth selection and computational acceleration, and the inference procedures for the performance of different tests. Section 3 presents the implementation of the methods in package **npregfast**. The package capabilities using a couple of real data examples are illustrated in Section 4 and lastly, Section 5 concludes with some remarks.

2. Statistical methodology

In many practical situations, the response variable, Y, depends on a continuous covariate, X. In such a regression framework, consideration might well be given to the nonparametric regression model

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

where m is a smooth unknown function and ε is the regression error with zero mean. The main advantage of using these type of models is the flexibility and the ease of interpretation of m.

A generalization of the "pure" model in (1) is the regression model with factor-by-curve interactions. In these type of models, the relationship between the response and the covariates can change depending on the levels of a categorical variable, F. The possibility of incorporating factor-by-curve interactions in nonparametric regression models has already been discussed by Hastie and Tibshirani (1990). Ruppert and Wand (1994) and Coull, Ruppert, and Wand (2001) also presented an algorithm based on penalized splines (P-splines), which would enable these types of interactions to be incorporated into these types of models. Recently, Cadarso-Suárez *et al.* (2006) and Roca-Pardiñas, Cadarso-Suárez, Nácher, and Acuña (2006) have successfully applied these interactions to estimate neuron firing rates.

Based on this, to study the possible effect of F on the response, the following nonparametric regression model including factor-by-curve interactions is considered

$$Y = f_0(X) + \begin{cases} f_1(X) + \varepsilon_1 & \text{if} \quad F = 1, \\ \cdots & \\ f_M(X) + \varepsilon_M & \text{if} \quad F = M, \end{cases} \tag{2}$$

where $\varepsilon_1, \ldots, \varepsilon_M$ are the zero-mean errors for each level of the factor, f_0 represents the global effect of X on the response, and f_l is the specific effect of X associated with the lth level of factor F. Note that under model (2), the regression curves $m_l(x) = \mathsf{E}(Y|X = x, F = l)$ are given by

$$m_l(X) = f_0(X) + f_l(X) \qquad \text{for} \quad l = 1, \ldots, M.$$

In order to prevent different combinations of $f_0, f_1, \ldots, f_M$ leading to the same model, the sum of the specific effects across the levels are assumed to be zero. That is to say, for each x, $\sum_{l=1}^{M} f_l(x) = 0$ is enforced. Note that this identifiability condition does not put any constraints on our model because it can be modified to conform to this condition.

In addition, when a factor-by-curve interaction is detected in model (2), it might be of interest to draw inferences about some critical points of curves (such as minima, maxima or inflection points), studying for this purpose the derivatives. In general, the critical point x_{0l} referring to the l level will be obtained from the derivative curve $m_l^r(x)$, for some r. Accordingly, we define this point, x_{0l}, for each l level, as

$$x_{0l} = \arg\max_x m_l^r(x). \tag{3}$$

The present section describes the estimation procedure for these types of models and for these critical points, based on the use of local polynomial kernel smoothers, and explains in detail the inference methods implemented in the package. It also shows the procedure used to select the bandwidth of the estimator and draws attention to the technique applied to speed up both the estimation and inference methods.

2.1. Estimation procedures

The factor-by-curve regression model in (2) is estimated using local polynomial kernel smoothers (Wand and Jones 1995; Fan and Gijbels 1996). Given a sample $\{(X_i, F_i, Y_i)\}_{i=1}^n$ of n independent and identically distributed (i.i.d.) observations, and considering observations in all the levels of F, the estimate of f_0 at a point x is given by $\hat{f}_0(x) = \hat{\alpha}_0(x)$, being $\hat{\alpha}_0(x)$ the first position of the vector $(\hat{\alpha}_0(x), \hat{\alpha}_1(x), \ldots, \hat{\alpha}_R(x))$ which is the minimizer of

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{r=0}^{R} \alpha_r(x) (X_i - x)^r \right\}^2 \cdot K\left(\frac{X_i - x}{h_0} \right), \tag{4}$$

where K is a kernel function (normally, a symmetric density), h_0 is the smoothing parameter or bandwidth and R is the degree of the polynomial. Moreover, the estimated rth ($r \leq R$) derivative of $f_0(x)$ is given by $\hat{f}_0^r(x) = r!\hat{\alpha}_r(x)$.

Once the estimation of f_0 is obtained, the estimate of f_l at a point x is given by $\hat{f}_l(x) = \hat{\alpha}_{0l}(x)$ (for $l = 1, \ldots, M$), being $\hat{\alpha}_{0l}$ the first position of the vector $(\hat{\alpha}_{0l}(x), \hat{\alpha}_{1l}(x), \ldots, \hat{\alpha}_{Rl}(x))$ which is the minimizer of

$$\sum_{i=1}^{n} \left\{ Y_i - \hat{f}_0(X_i) - \sum_{r=0}^{R} \alpha_{rl}(x)(X_i - x)^r \right\}^2 \cdot K\left(\frac{X_i - x}{h_l}\right) I_{\{F_i = l\}}, \tag{5}$$

where h_l is the bandwidth used to obtain $\hat{f}_l$ and I the indicator function. Analogously, the estimated rth ($r \leq R$) derivative of $f_l(x)$ is given by $\hat{f}_l^r(x) = r!\hat{\alpha}_{rl}(x)$.

Note that the obtained estimates do not necessarily meet the imposed identifiability condition. To do so, the following procedure is used. For each x, calculate the mean of the specific effects of each level, $S(x) = M^{-1}\sum_{l=1}^{M} \hat{f}_l(x)$, and replace the original $\hat{f}(x)$ and $\hat{f}_l(x)$ by $\hat{f}_l(x) - S(x)$ and $\hat{f}_0(x) + S(x)$, respectively.

Clearly, the estimated curves for each level at point x are given by $\hat{m}_l(x) = \hat{f}_0(x) + \hat{f}_l(x)$, for $l = 1, \ldots, M$, and the estimated rth derivative of $m_l(x)$ is given by $\hat{m}_l^r(x) = \hat{f}_0^r(x) + \hat{f}_l^r(x)$.

Finally, a natural estimator of the critical point x_{0l} (3) can be obtained as the maximizer of

$$\hat{m}_l^r(k_1), \ldots, \hat{m}_l^r(k_N)$$

with $k_1, \ldots, k_N$ being a grid of N equidistant points in a range of X values.

Note that the proposed methodology makes sense when the support of X is the same for all the levels and it is also a closed and bounded interval.

2.2. Inference procedures

The procedures implemented in this package enable us to test two hypotheses. The first one is a global test which assumes the hypothesis of equality of the M regression functions (or derivatives) and the second one is a local test that enables us to test the hypothesis that, among the levels of a given factor, the critical points are equal.

Global test

Here we expose a procedure to test the following null hypothesis based on the model in (2):

$$H_0^r : m_1^r(\cdot) = \cdots = m_M^r(\cdot) \tag{6}$$

versus the general alternative

$$H_1^r : m_i^r(\cdot) \neq m_j^r(\cdot) \qquad \text{for some} \quad i, j \in \{1, \ldots, M\}.$$

It should be noted that the previous hypothesis is equivalent to $f_1^r(\cdot) = \cdots = f_M^r(\cdot) = 0$, and therefore $f_l(x) = \sum_{j=0}^{r-1} a_{jl}x^j$ will be a polynomial of degree $r-1$ for $l = 1, \ldots, M$.[1]

[1]Let us assume that $f^r(x) = g^r(x)$ for all x. Let $h(x) = f(x) - g(x)$. Hence, $h^r(x) = f^r(x) - g^r(x) = 0$. By applying Taylor's theorem to the function h up to order r, and taking into account that the derivatives of h of order higher or equal than r are zero, we obtain

$$h(x) = f(x) - g(x) = h(0) + h^1(0)x + \frac{h^2(0)}{2!}x^2 + \cdots + \frac{h^{r-1}(0)}{(r-1)!}x^{r-1},$$

which shows that $h(x)$ is a polynomial of degree $r-1$.

Accordingly, the null regression model is given by

$$Y = f_0(X) + \begin{cases} \sum_{j=0}^{r-1} a_{j1} X^j + \varepsilon_1 & \text{if} \quad F = 1, \\ \cdots & \\ \sum_{j=0}^{r-1} a_{jM} X^j + \varepsilon_M & \text{if} \quad F = M, \end{cases} \tag{7}$$

and the regression curves m_l are given by $m_l(X) = f_0(X) + \sum_{j=0}^{r-1} a_{jl} X^j$. Note that, in the expression (7), we have abused notation slightly. In fact, if $r = 0$ we are actually referring to the null model $Y = f_0(X) + \varepsilon$.

To test H_0^r, we propose the use of the following test statistic based on direct nonparametric estimates of f_l^r curves considering the L_1 norm

$$T = \sum_{l=1}^{M} \frac{n_l}{n} \sum_{i=1}^{n} |\hat{f}_l^r(X_i) I_{\{F_i = l\}}|,$$

being $n_l = \sum_{i=1}^{n} I_{\{F_i = l\}}$. For a detailed simulation study comparing other test statistics see Sestelo (2013).

Note that if H_0^r holds, the value of T should be close to zero. The test rule based on T consists of rejecting the null hypothesis if T is larger than its $(1-\alpha)$-percentile obtained under H_0. To approximate the distributions of the test statistic resampling methods such as the bootstrap introduced by Efron (1979) (see also Efron and Tibshirani 1993; Härdle and Mammen 1993; Kauermann and Opsomer 2003) can be applied instead. Here we use the wild bootstrap (Wu 1986; Liu 1988; Mammen 1993) because this method is valid also for heteroscedastic models where the variance of the error is a function of the covariate. The testing procedure used here involves the following steps:

Step 1. Compute the value of the test statistic, T, in the sample as explained above.

Step 2. Estimate the null regression model in (7). For this purpose, estimate $f_0(X_i)$ as we mentioned in the estimation procedure in Section 2.1. Calculate $Y_i^l = Y_i - \hat{f}_0(X_i)$ and with that fit the polynomial using least squares for each level. Obtain the pilot estimates for $i = 1, \ldots, n$,

$$\hat{m}_{F_i}(X_i) = \hat{f}_0(X_i) + \sum_{j=0}^{r-1} \hat{a}_{jF_i} X_i^j.$$

Step 3. For $b = 1, \ldots, B$, generate bootstrap samples $\left\{\left(X_i, F_i, Y_i^{\bullet b}\right)\right\}_{i=1}^{n}$ with $Y_i^{\bullet b} = \hat{m}_{F_i}(X_i) + \varepsilon_i^{\bullet b}$, and $\varepsilon_i^{\bullet b}$ being

$$\varepsilon_i^{\bullet b} = \begin{cases} \hat{\varepsilon}_i \cdot \frac{(1-\sqrt{5})}{2} & \text{with probability } p = \frac{5+\sqrt{5}}{10}, \\ \hat{\varepsilon}_i \cdot \frac{(1+\sqrt{5})}{2} & \text{with probability } p = \frac{5-\sqrt{5}}{10}, \end{cases}$$

where $\hat{\varepsilon}_i = Y_i - \hat{m}_{F_i}(X_i)$ are the residuals under H_0, and compute $T^{\bullet b}$ as in Step 1.

Finally, the decision rule consists of rejecting the null hypothesis if $T > T^{1-\alpha}$, where $T^{1-\alpha}$ is the empirical $(1-\alpha)$-percentile of values $T^{\bullet b} (b = 1, \ldots, B)$ previously obtained.

Local test

If the previous test is statistical significant and the equality of the m_l^r curves ($l = 1, \ldots, M$) is thus rejected, testing the null hypothesis of equality of critical points becomes of interest. Note that it is possible for these points to be equal, even if the curves and/or their derivatives are different. For instance, taking into account the maxima of the first derivatives, interest lies in testing the following null hypothesis

$$H_0 : x_{01} = \cdots = x_{0M}$$

versus the general alternative

$$H_1 : x_{0i} \neq x_{0j} \qquad \text{for some} \quad i, j \in \{1, \ldots, M\}.$$

The above null hypothesis is true if $d = x_{0j} - x_{0k} = 0$ where

$$(j, k) = \underset{\substack{(l,m) \\ \{1 \leq l < m \leq M\}}}{\arg\max} \; |x_{0l} - x_{0m}|,$$

otherwise H_0 is false. It is important to highlight the fact that, in practice, the true x_{0j} are not known, and consequently neither is d, so an estimate $\hat{d} = \hat{x}_{0j} - \hat{x}_{0k}$ is used, where, in general, $\hat{x}_{0l}$ are the estimates of x_{0l} based on the estimated curves $\hat{m}_l$.

Needless to say, since $\hat{d}$ is only an estimate of the true d, the sampling uncertainty of these estimates needs to be taking into account. Hence, a confidence interval may be created for d at a specific level of confidence. Based on this, the null hypothesis is rejected if zero is not contained in the interval.

The steps for construction of the bootstrap confidence interval for the true d are the following:

Step 1. From the sample data $\{(X_i, F_i, Y_i)\}_{i=1}^n$, obtain the estimates for $i = 1, \ldots, n$

$$\hat{m}_{F_i}(X_i) = \hat{f}_0(X_i) + \hat{f}_{F_i}(X_i)$$

based on the general model in (2), obtain the estimates of x_{0l} based on (3) and then retrieve the $\hat{d}$ value.

Step 2. For $b = 1, \ldots, B$, generate bootstrap samples $\left\{\left(X_i, F_i, Y_i^{\bullet b}\right)\right\}_{i=1}^n$ as in Step 3 of the algorithm for the global test presented earlier, though, in this case, using the residuals of the general model in (2), $\hat{\varepsilon}_i = Y_i - \hat{m}_{F_i}(X_i)$, and compute $d^{\bullet b}$ as in Step 1.

Finally, the limits for the $100(1 - \alpha)\%$ percentile confidence interval of d are given by

$$I = \left(\hat{d}^{\alpha/2}, \hat{d}^{1-\alpha/2}\right),$$

where $\hat{d}^p$ represents the p-percentile of $\hat{d}^{\bullet 1}, \ldots, \hat{d}^{\bullet B}$.

2.3. More technical details

It is well known that the nonparametric estimates $\hat{m}_l^r(X)$ greatly depend on the bandwidths $h_0, h_1, \ldots, h_M$ used in the kernel-based algorithm for the estimation of the partial functions

$f_0, f_1, \ldots, f_M$. Various methods for an optimal selection have been suggested, such as generalized cross-validation (GCV; Golub, Heath, and Wahba 1979) or plug-in methods (see e.g., Ruppert, Sheather, and Wand 1995). See Wand and Jones (1995) for a good overview of this topic. However, optimal bandwidth selection is still a challenging problem.

As a practical solution, in Equation 4 of the estimation algorithm, the bandwidth h_0 is automatically selected by minimizing the following cross-validation criterion:

$$CV_0(h) = \sum_{i=1}^{n} \left(Y_i - \hat{f}_0^{(-i)}(X_i) \right)^2, \tag{8}$$

where $\hat{f}_0^{(-i)}(X)$ indicates the fit at X, leaving out the ith data point based on the smoothing parameter h_0. Likewise, the bandwidths h_l $(l = 1, \ldots, M)$ of Equation 5 are selected by minimizing

$$CV_l(h) = \sum_{i=1}^{n} I_{\{F_i = l\}} \left(Y_i - \hat{f}_0(X_i) - \hat{f}_l^{(-i)}(X_i) \right)^2, \tag{9}$$

where $\hat{f}_l^{(-i)}(X)$ indicates the fit at X, leaving out the ith data point based on the smoothing parameter h_l.

Bootstrap resampling techniques are time-consuming processes because it is necessary to estimate the model many times. Moreover, the use of the cross-validation technique for the choice of the bandwidths implies a high computational cost, because it is necessary to repeat the estimation operations several times to select the optimal bandwidths. Consequently, recourse to some computational acceleration technique is fundamental to ensure that the problem can be addressed adequately in practical situations. Thus, we use binning techniques to speed up the process. A detailed explanation of this technique can be found in Fan and Marron (1994).

3. Overview of the package npregfast

The **npregfast** package contains a set of functions for estimating nonparametric models, obtaining first and second derivatives, critical points, etc., as well as different tests for drawing inferences about several features of these models. In view of the high cost entailed in these methodologies and in order to maximize computational efficiency, the actual version of the package is carried out using compiled Fortran. The functions within **npregfast** are briefly described in Table 1.

The package is designed along lines similar to those of other R regression packages. Hence, the main function of the package is `frfast` which, by default, fits a nonparametric regression model based on local polynomial kernel smoothers. The arguments of this function are shown in Table 2. Note that through the argument `formula` users can decide to fit a model taking into account the interaction or not, and by means of the argument `smooth` it is possible to select the type of smoother: kernel or splines. Numerical and graphical summaries of the fitted object can be obtained by using the `print`, `summary`, `plot` and `autplot` methods implemented for '`frfast`' objects (arguments of the latter function are shown in Table 3). Another of these methods is available for the `predict` function, which takes a fitted model of the '`frfast`' class and, given a new data set of values of the covariate, produces predictions.

As mentioned above, this package can be used to fit models taking into account factor-by-curve interactions. In this framework, it will be necessary to ascertain if the factor produces an effect

Function	Description
`frfast`	Main function for fitting regression models and obtaining the different outputs (model estimates, first and second derivatives).
`summary`	Method of the generic summary function for '`frfast`' objects.
`autoplot`	Visualization of '`frfast`' objects with **ggplot2** (Wickham 2009) graphics. Provides the plots for model estimates and their 95% pointwise confidence intervals based on bootstrap techniques. Additionally, with the `diffwith` argument it is possible to draw the differences between two factor levels.
`plot`	Visualization of '`frfast`' objects with base graphics. Provides the plots for model estimates and their 95% pointwise confidence intervals based on bootstrap techniques. Additionally, with the `diffwith` argument it is possible to draw the differences between two factor levels.
`predict`	Takes a '`frfast`' object produced by `frfast()` and, given a new set of values for the model covariate, produces predictions.
`critical`	Provides a table with the value of the covariate `x` (with 95% confidence interval) that maximizes the initial estimation, that maximizes the first derivative and where the second derivative equals zero.
`criticaldiff`	Provides a table with the 95% confidence interval for the differences between the estimation of the `critical` function, for every two levels.
`globaltest`	Function for testing the equality of the curves specific to each level.
`localtest`	Function for testing the equality of the critical points estimated from the respective level-specific curves.
`allotest`	Function for testing the null hypothesis of an allometric model *versus* a general hypothesis where the effect of the covariate on the response is flexible and unknown.
`runExample`	Launch a Shiny app that shows a demo of what can be done with the package.

Table 1: Summary of functions in the **npregfast** package.

on the response and thus, that there is an interaction or, in contrast, the estimated regression curves are equal. To this end, the package provides the `globaltest` function which answers this question through a bootstrap-based test. If the factor results to be statistically significant, then the use of the `diffwith` argument of the `autoplot` method for '`frfast`' objects (or of its base graphics version, the `plot` method for '`frfast`' objects) enables the user to obtain a graphical representation that shows the differences between the estimated curves (estimate, first or second derivative) for any set of two levels of the factor. Additionally, the function `critical` allows to obtain the value of the covariate that maximizes the estimate and first derivative of the function and the value of the covariate where the second derivative equals zero, for each of these levels. Again, to test if these estimated points are equal for all levels, the package provides the `localtest` function. Note that, to compare these points between any set of two levels, a confidence interval for the difference can be obtained by applying the `criticaldiff` function. It should be noted that both smoothing methods (kernel and splines) are available options to test the equality of the M curves specific to each level, or to test the equality of critical points (i.e., for the `globaltest` and `localtest` functions).

Arguments	Description
`formula`	An object of class 'formula'; a symbolic description of the model to be fitted.
`data`	A data frame or matrix containing the model response variable and covariates required by the `formula`.
`na.action`	A function which indicates what should happen when the data contains `NA`s. The default is `"na.omit"`.
`model`	Type of model used: `model = "np"` for the nonparametric regression model, `model = "allo"` for the allometric model.
`smooth`	Type of smoother used: `smooth = "kernel"` for kernel smoothers and `smooth = "splines"` for splines using the **mgcv** package (Wood 2017).
`h0`	The kernel bandwidth smoothing parameter for the global effect. By default, cross-validation is used to obtain the bandwidth.
`h`	The kernel bandwidth smoothing parameter for the partial effects.
`nh`	Integer number of equally-spaced bandwidth in which the `h` is discretized, to speed up computation.
`weights`	Prior weights on the data.
`kernel`	A character string specifying the desired kernel. Defaults to `kernel = "epanech"`, where the Epanechnikov density function kernel will be used. Also, several types of kernel functions can be used: triangular and Gaussian density function, with `"triang"` and `"gaussian"`, respectively.
`p`	Polynomial degree to be used. Its value must be the value of derivative + 1. The default value is 3, returning the estimation, first and second derivative.
`kbin`	Number of binning nodes over which the function is to be estimated.
`nboot`	Number of bootstrap repeats. Defaults to 500 bootstrap repeats. The wild bootstrap is used when `model = "np"` and the simple bootstrap when `model = "allo"`.
`rankl`	Number or vector specifying the minimum value for an interval at which to search for the `x` value that maximizes the estimate, and first or second derivative (for each level). The default is the minimum data value.
`ranku`	Number or vector specifying the maximum value for an interval at which to search for the `x` value that maximizes the estimate, and first or second derivative (for each level). The default is the maximum data value.
`seed`	Seed to be used in the bootstrap procedure.
`cluster`	A logical value. If `TRUE` (default), the bootstrap procedure is parallelized (only for `smooth = "splines"`). Note that there are cases (e.g., a low number of bootstrap repetitions) that R will gain in performance through serial computation. R takes time to distribute tasks across the processors and also it will need time for binding them all together later on. Therefore, if the time for distributing and gathering pieces together is greater than the time needed for single-thread computing, it is not worth to parallelize.
`ncores`	An integer value specifying the number of cores to be used in the parallelized procedure. If `NULL` (default), the number of cores to be used is equal to the number of cores of the machine – 1.

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

Arguments	Description
`object`	'`frfast`' object.
`fac`	Factor level to be taken into account in the plot. By default, `NULL`.
`der`	Number which determines any inference process. By default, `NULL`. If this term is `0`, the plot shows the initial estimate. If it is `1` or `2`, it is designed for the first or second derivative, respectively.
`diffwith`	Factor level used for drawing the differences with respect to the level specified in the `fac` argument. By default, `NULL`. The differences are computed for the rth derivative specified in the `der` argument.
`points`	Draw the original data into the plot. By default, `TRUE`.
`xlab`	A title for the `x` axis.
`ylab`	A title for the `y` axis.
`ylim`	The `y` limits of the plot.
`main`	An overall title for the plot.
`col`	A specification for the default plotting color.
`CIcol`	A specification for the default confidence intervals plotting color (for the fill).
`CIlinecol`	A specification for the default confidence intervals plotting color (for the edge).
`pcol`	A specification for the point color.
`abline`	Draw an horizontal line into the plot of the second derivative of the model.
`ablinecol`	The color to be used for `abline`.
`lty`	The line type. Line types can either be specified as an integer (0 = blank, 1 = solid (default), 2 = dashed, 3 = dotted, 4 = dotdash, 5 = longdash, 6 = twodash) or as one of the character strings `"blank"`, `"solid"`, `"dashed"`, `"dotted"`, `"dotdash"`, `"longdash"`, or `"twodash"`, where `"blank"` uses "invisible lines" (i.e., does not draw them). See details in `?par`.
`CIlty`	The line type for confidence intervals. Line types can either be specified as an integer (0 = blank, 1 = solid (default), 2 = dashed, 3 = dotted, 4 = dotdash, 5 = longdash, 6 = twodash) or as one of the character strings `"blank"`, `"solid"`, `"dashed"`, `"dotted"`, `"dotdash"`, `"longdash"`, or `"twodash"`, where `"blank"` uses "invisible lines" (i.e., does not draw them). See details in `?par`.
`lwd`	The line width, a positive number, defaulting to 1. See details in `?par`.
`CIlwd`	The line width for confidence intervals, a positive number, defaulting to 1.
`cex`	A numerical value giving the amount by which plotting symbols should be magnified relative to the default. See details in `?par`.
`alpha`	Alpha transparency for overlapping elements expressed as a fraction between 0 (complete transparency) and 1 (complete opacity).
`...`	Other options.

Table 3: Summary of the arguments of the `autoplot` method for '`frfast`' objects.

4. npregfast in practice

It is now time to outline the implemented functions of our package in detail, and illustrate these with two real data sets related to the life sciences. The first one is connected with the biology and management of an aquatic living resource and the second one is related to medical data, particularly, the age and height measurements of children.

4.1. Relative growth curves for barnacles

The **npregfast** package includes a data set called `barnacle` with measurements of rostro-carinal length and dry weight of barnacles from the Atlantic coast of Galicia (Spain), split by a categorical variable indicating the site of harvest (Sestelo and Roca-Pardiñas 2011). The usage of the package is illustrated by constructing the relative growth curves for this species and determining the ideal size of capture of the stalked barnacle, *Pollicipes pollicipes* (Gmelin, 1789). The commercial interest of this crustacean resides in their muscular peduncle, the edible stalk of the barnacle, which commands high prices on the market (Goldberg 1984). In Spain and Portugal, where harvesting of *P. pollicipes* is the highest, the phenomenon of overfishing has affected this species to differing degrees (Bernard 1988; Cardoso and Yule 1995; Cruz 2000; Molares and Freire 2003). Because of the economic importance of this barnacle in several countries, we appreciate to deepen our knowledge about it. Accordingly, the main goal of this data set is to illustrate the use of the R package to analyze the relationship between the gain in weight and length of barnacles.

Each line of the data set represents the information from one specimen under study. The `DW` variable denotes the dry weight of the individuals in grams, the `RC` variable is the rostro-carinal length in millimeters – the variable that best represents the growth of the species (Cruz 1993, 2000) – and the categorical variable `F` indicates the site where the specimens were collected, Punta Lens (`lens`) and Punta de la Barca (`barca`). An excerpt of the data set is shown below:

```
R> library("npregfast")
R> head(barnacle)[1:3, ]
```

```
    DW   RC     F
1 0.14  9.5 barca
2 0.00  2.4 barca
3 0.42 13.1 barca
```

To estimate the length-weight relationship of this species we first consider a nonparametric regression model without interaction. The polynomial degree was fixed to 2 based on the idea of using later the first derivative.

```
R> mwo <- frfast(DW ~ RC, data = barnacle, p = 2, seed = 130853)
R> mwo
```

```
Call:
frfast(formula = DW ~ RC, data = barnacle, p = 2, seed = 130853)


*******************************************
Nonparametric Model
*******************************************

Number of Observations: 2000
```

```
Number of Bootstrap Repeats: 500

Type of Nonparametric Smoother: kernel

Bandwidth: 0.21

Kernel Function: Epanechnikov
```

Note that, by default, the function `frfast` fits a flexible model using local polynomial kernel smoothers where the bandwidth is selected by cross-validation. The bootstrap procedure is used to obtain the 95% confidence interval for the estimation.

The graphical representation of the fit is obtained using the `autoplot` function. Figure 1 (left panel) shows the estimated curve (solid line) for the overall study with the pointwise confidence interval (shaded area). This curve shows the way in which individuals' size increased as their weight increased. The length-weight relationship is seen to be an increasing function in almost the complete range of values; only the final section of the curve seems to stabilize to a horizontal line. The way in which the gain weight occurs can be obtained by means of the first derivative (Figure 1, right panel). The speed of growth (the increase in weight per unit of `RC`), rather than constantly increasing, displayed a maximum at a specific size, after which it began to decrease. Both plots can be plotted jointly with the following input commands:

```
R> library("gridExtra")
R> der0 <- autoplot(mwo, der = 0)
R> der1 <- autoplot(mwo, der = 1)
R> grid.arrange(der0, der1, nrow = 1, ncol = 2)
```

In biological studies, and specifically in population dynamics and stock assessment, it is relevant to ascertain whether this length-weight relationship remains constant across sites and was not altered by any possible local variability in the growth of this species. Therefore, we now intend to estimate the model including the factor-by-curve interaction. This can be obtained with the following code:

```
R> mwi <- frfast(DW ~ RC:F, data = barnacle, p = 2, seed = 130853)
R> summary(mwi)

Call:
frfast(formula = DW ~ RC:F, data = barnacle, p = 2, seed = 130853)

******************************************
Nonparametric Model
******************************************
Type of nonparametric smoother:  kernel
Kernel: Epanechnikov
Bandwidth: 0.21 0.31 1.00
Polynomial degree: 2
Number of bootstrap repeats: 500
Number of binning nodes 100
```

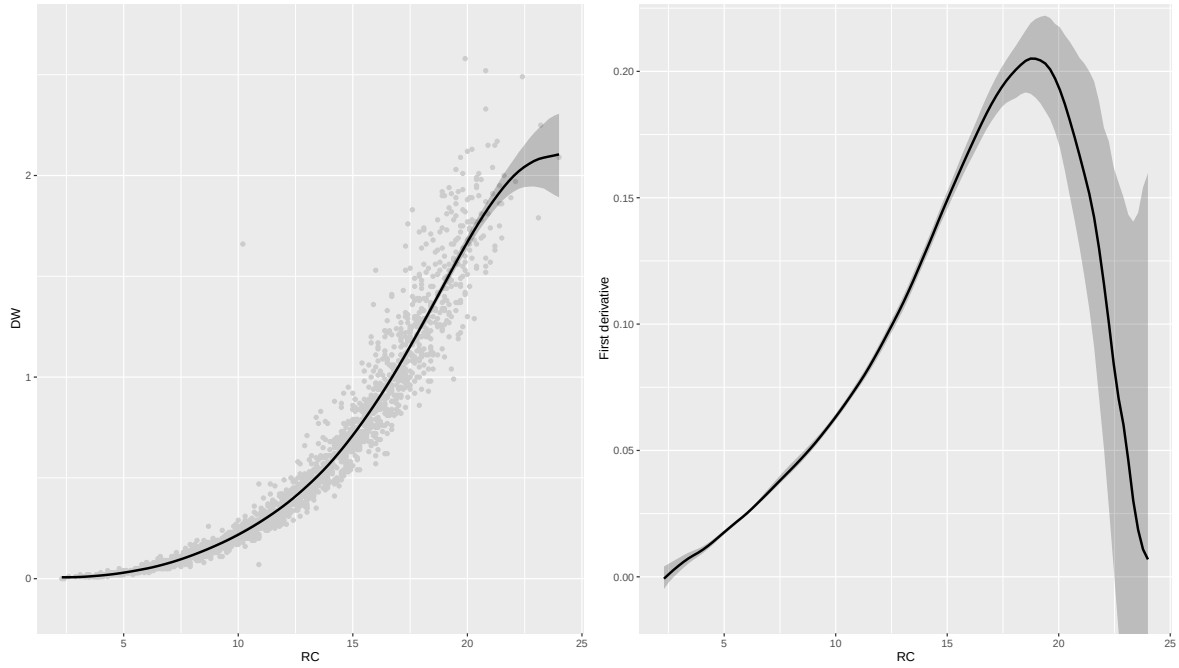

Figure 1: Regression curve (left) and first derivative (right) with pointwise bootstrap-based 95% confidence intervals (shaded area) for dry weight (`DW`) and rostro-carinal length (`RC`) (overall study).

```
The number of data is:  2000
The factor's levels are:  barca lens
The number of data for the level barca is: 1000
The number of data for the level lens is: 1000

Summaries for the response variable (for each level):
Level barca :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1300  0.4100  0.5437  0.8425  2.2500

Level lens :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1400  0.4350  0.5974  0.9500  2.5800
```

The `summary` method returns a numerical summary of the fit where it is possible to observe the kernel used, the global and partial bandwidths obtained by cross-validation, the number of bootstrap repeats used for obtaining the confidence interval for the estimation, the number of binning nodes, the sample size by levels of the factor and a small summary for the response by levels.

The fit can be again visualized by means of the generic function `autoplot`. As in the previous case, it is possible to represent both the estimation and first derivative with the `der` argument.

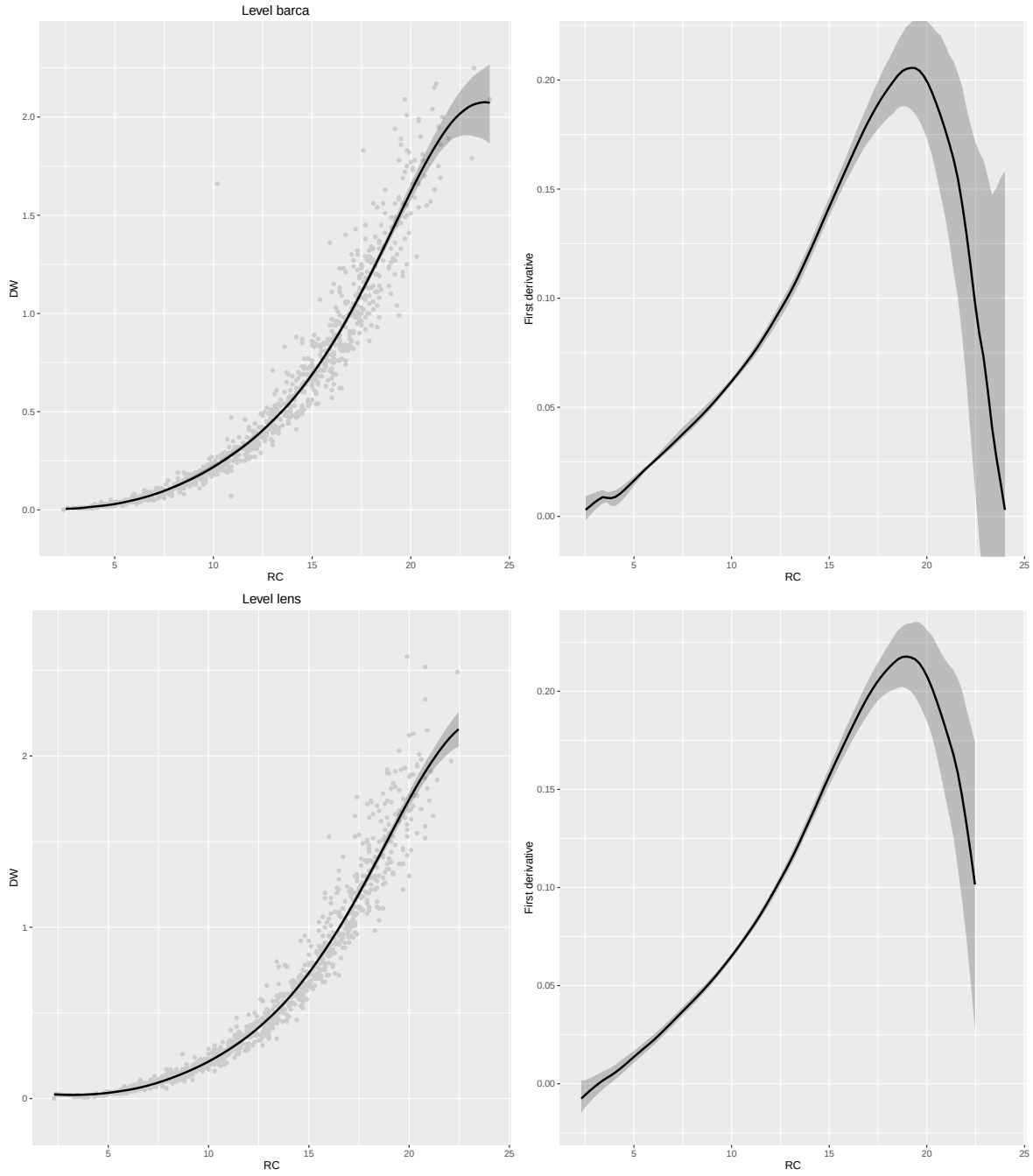

Figure 2: Regression curve (left) and first derivative (right) with bootstrap-based 95% confidence intervals (shared area) for dry weight (DW) and rostro-carinal length (RC) for barnacles from Punta de la Barca (upper panel) and for barnacles from Punta Lens (lower panel).

Additionally, the selection of the factor level is obtained by the `fac` argument. Figure 2 shows the estimated length-weight relationship for the barnacles of the two sites of Galicia, Punta de la Barca and Punta Lens.

```
R> library("ggplot2")
```

```
R> der01 <- autoplot(mwi, der = 0, fac = "barca")
R> der11 <- autoplot(mwi, der = 1, fac = "barca") + ggtitle("")
R> der02 <- autoplot(mwi, der = 0, fac = "lens")
R> der12 <- autoplot(mwi, der = 1, fac = "lens") + ggtitle("")
R> grid.arrange(der01, der11, der02, der12, nrow = 2, ncol = 2)
```

The question that now arises is if the estimated curves for each level are identical and thus indicate that there is no need to include the interaction in the model or, by contrast, the barnacles show a difference in relative growth and the factor really produces an effect on the response. To this end, the bootstrap-based test proposed before for testing the equality of the M regression functions (or derivatives) was applied using the `globaltest` function.

```
R> globaltest(DW ~ RC:F, data = barnacle, p = 2, seed = 130853, der = 0)

  Statistic pvalue Decision
1  5.194234      0 Rejected

R> globaltest(DW ~ RC:F, data = barnacle, p = 2, seed = 130853, der = 1)

  Statistic pvalue Decision
1  1.221476  0.006 Rejected
```

Taking into account the results obtained, we can conclude that both the estimates and derivatives are not equal between levels. This is also observed from the graphical representations. The test concludes that the factor site produces an effect on the response.

Finally, an important issue related with any species that is subject to exploitation is the establishment of limits on size. The estimation of adequate catch sizes for commercial marine invertebrates includes biological aspects such as individual size at sexual maturation, growth rate and length-weight relationship but also the yield in weight from the fishery (Sparre and Venema 1997). According to this, the point that maximizes the first derivative of the regression curve must be determined. This critical point would ensure high commercial yield while simultaneously guaranteeing the regeneration and conservation of the population. To obtain these points for both sites of harvest, the `critical` function can be applied.

```
R> critical(mwi, der = 1)

            Critical      Lwr      Upr
Level barca  19.1995 18.56957 21.00946
Level lens   19.0040 18.45064 19.77458
```

The estimated critical points with their 95% confidence intervals seem to be similar between the two sites of study. To ascertain the truth of this affirmation, the `localtest` function for testing the equality of critical points was applied. In this case, one tests whether the points that maximize the first derivatives of the curves are equal.

```
R> localtest(DW ~ RC:F, data = barnacle, p = 2, seed = 130853, der = 1)
```

```
       d     Lwr    Upr Decision
1 0.1955 -0.1521 1.1893 Accepted
```

According to the obtained confidence interval it is possible to conclude that, although the effects of the size (RC) on the weight (DW) depend on the location (F) and consequently the curves of the relative growth and their derivatives are different for each level, there is not a statistically significant difference between the estimated sizes.

4.2. Spurt in child growth

We decided to also show the capabilities of package **npregfast** with another data set, which contains the age and height measurements of 2500 children aged 5 to 19 years, split by sex (1292 females and 1208 males). The usage of the package is illustrated by constructing growth curves for school-aged children and adolescents and also by analyzing possible differences in the growth of boys and girls. Other studies of this type can be obtained from http://www.who.int/childgrowth/en/. Finally, note that we applied in this section the two smoothers implemented in the package (kernel and splines) in order to compare the possible differences in the results. Below is an excerpt from the data frame of the data set used:

```
R> head(children)[1:3, ]
```

```
     sex height   age
1   male 150.77 13.25
2 female 170.59 14.17
3 female 167.31 15.17
```

Each line represents the information from one individual under study. The categorical variable sex indicates the individual's gender (male or female), the age variable corresponds to age in years, and height is measured in centimeters. To estimate the growth of the children overall, we firstly consider a nonparametric model without interaction.

```
R> mwo2k <- frfast(height ~ age, data = children, p = 2, seed = 130853)
R> mwo2k
```

```
Call:
frfast(formula = height ~ age, data = children, p = 2, seed = 130853)

*******************************************
Nonparametric Model
*******************************************

Number of Observations: 2500

Number of Bootstrap Repeats: 500

Type of Nonparametric Smoother: kernel
```

```
Bandwidth: 0.28

Kernel Function: Epanechnikov
```

One can obtain the results from the previous model based on spline smoothing. The spline-based model is obtained using argument `smooth = "splines"` (kernel is the default smoother) of the function `frfast`.

```
R> mwo2s <- frfast(height ~ s(age), data = children, seed = 130853,
+    smooth = "splines")
R> mwo2s

Call:
frfast(formula = height ~ age, data = children, smooth = "splines",
    seed = 130853)

******************************************
Nonparametric Model
******************************************

Number of Observations: 2500

Number of Bootstrap Repeats: 500

Type of Nonparametric Smoother: splines
```

The graphical representation of the fitted models can easily be obtained. Figure 3 plots the estimated curves obtained by means of the two smoothers with their 95% pointwise confidence intervals. As expected, in both cases, children's height rises with the increase in years of life until they reach a specific age; and thereafter their heights remain more or less constant. We can also observe that estimates obtained using the two smoothers are very similar. This plot can be obtained by using the following commands:

```
R> der0k <- autoplot(mwo2k, der = 0)
R> der0s <- autoplot(mwo2s, der = 0)
R> grid.arrange(der0k, der0s, nrow = 1, ncol = 2)
```

A common issue is to compare the growth between boys and girls. With this in mind, we fit a model taking into account the interaction. Again, we estimate the proposed model using both smoothers. It is worth mentioning that argument `formula` of the main function `frfast` must be specified in a different manner depending on the chosen smoother.

```
R> mwi2k <- frfast(height ~ age:sex, data = children, p = 2, seed = 130853)
R> mwi2s <- frfast(height ~ s(age, by = sex), data = children, seed = 130853,
+    smooth = "splines")
```

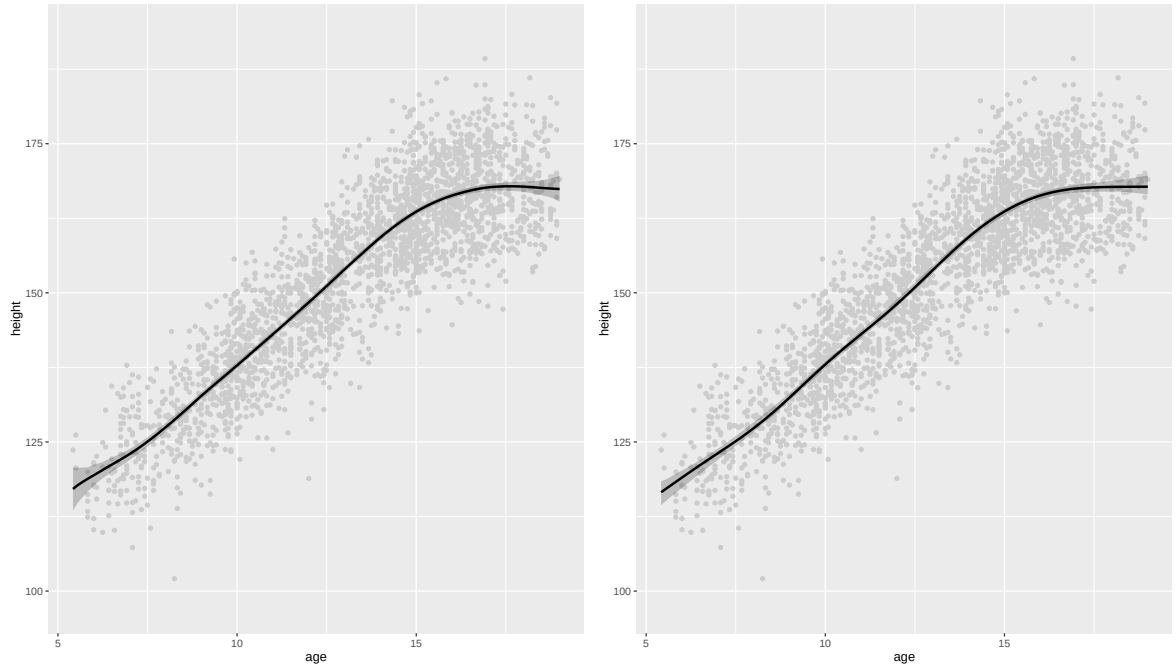

Figure 3: Regression curve (solid lines) with bootstrap-based 95% pointwise confidence intervals (shaded area) for children's height and age (overall study). Left panel: kernel smoothers. Right panel: splines.

The estimated curves for each gender can again be obtained with the function `autoplot`. Figures 4 and 5 show the estimated curves for males and females together with their first derivatives using both smoothers. The obtained estimates are similar, though they seem to be slightly smoother when using spline smoothing.

```
R> derk1 <- lapply(0:1, function(x) autoplot(mwi2k, der = x, fac = "male"))
R> derk2 <- lapply(0:1, function(x) autoplot(mwi2k, der = x,
+    fac = "female"))
R> grid.arrange(grobs = c(derk1, derk2), nrow = 2, ncol = 2)
R> ders1 <- lapply(0:1, function(x) autoplot(mwi2s, der = x, fac = "male"))
R> ders2 <- lapply(0:1, function(x) autoplot(mwi2s, der = x,
+    fac = "female"))
R> grid.arrange(grobs = c(ders1, ders2), nrow = 2, ncol = 2)
```

It is now time to asses if the factor really produces an effect on the response. To this end, we apply the bootstrap-based test implemented in `globaltest()`. Judging by the function output, the results would appear to suggest that the factor, `sex`, produces a real influence on the children's growth. This can also be observed from the graphical representation. Similarly, it can be concluded that the derivatives of these curves are different between levels. Note that the selection of the smoother does not change the hypothesis test conclusion.

```
R> globaltest(height ~ age:sex, data = children, p = 2, seed = 130853,
+    der = 0)
```

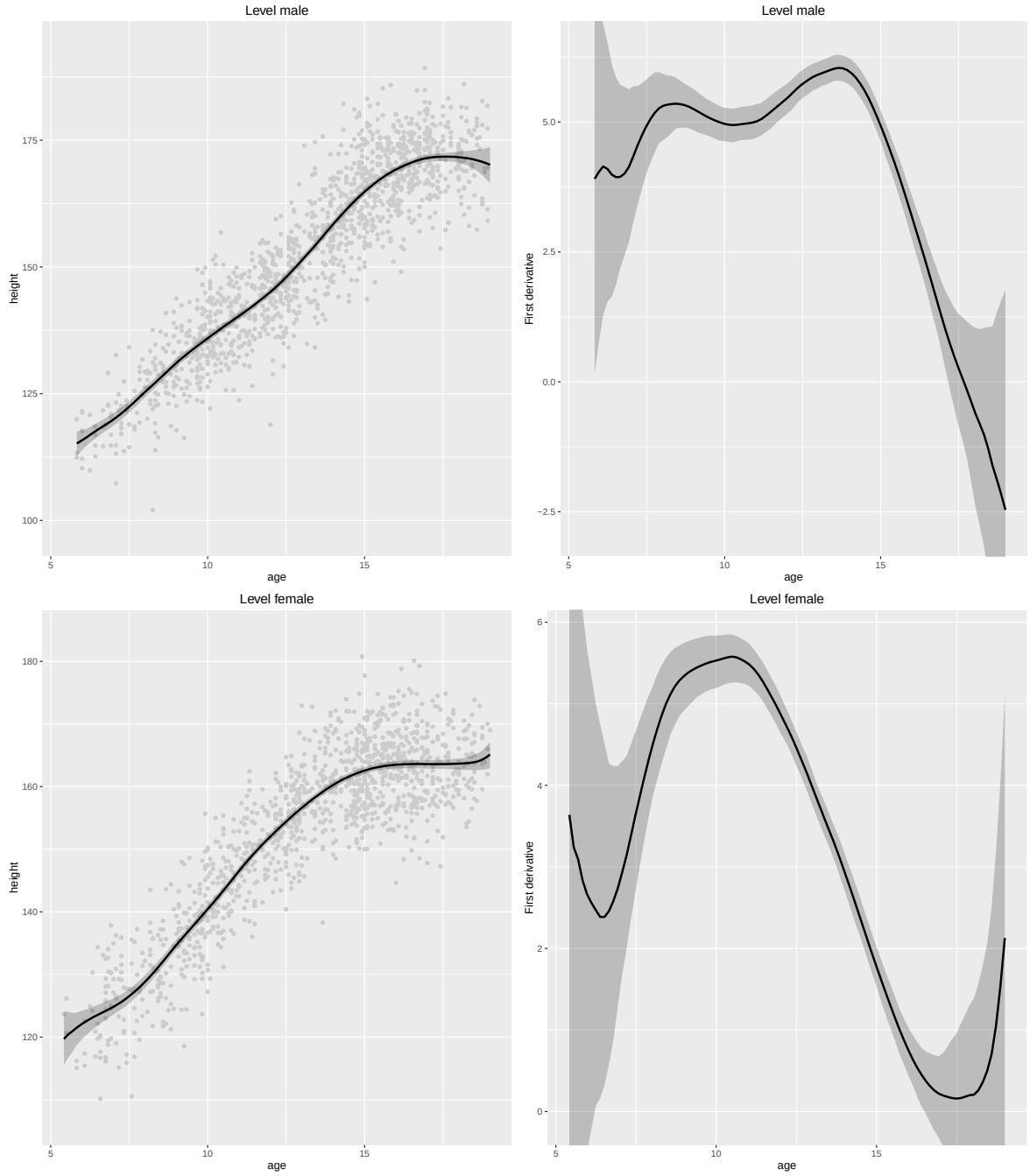

Figure 4: Regression curve and first derivative (solid lines) with bootstrap-based 95% pointwise confidence intervals (shaded area) for height and age of males (first row) and females (second row) using kernel smoothers.

```
  Statistic pvalue Decision
1  517.4986      0 Rejected
```

```
R> globaltest(height ~ s(age, by = sex), data = children, seed = 130853,
+    der = 0, smooth = "splines")
```

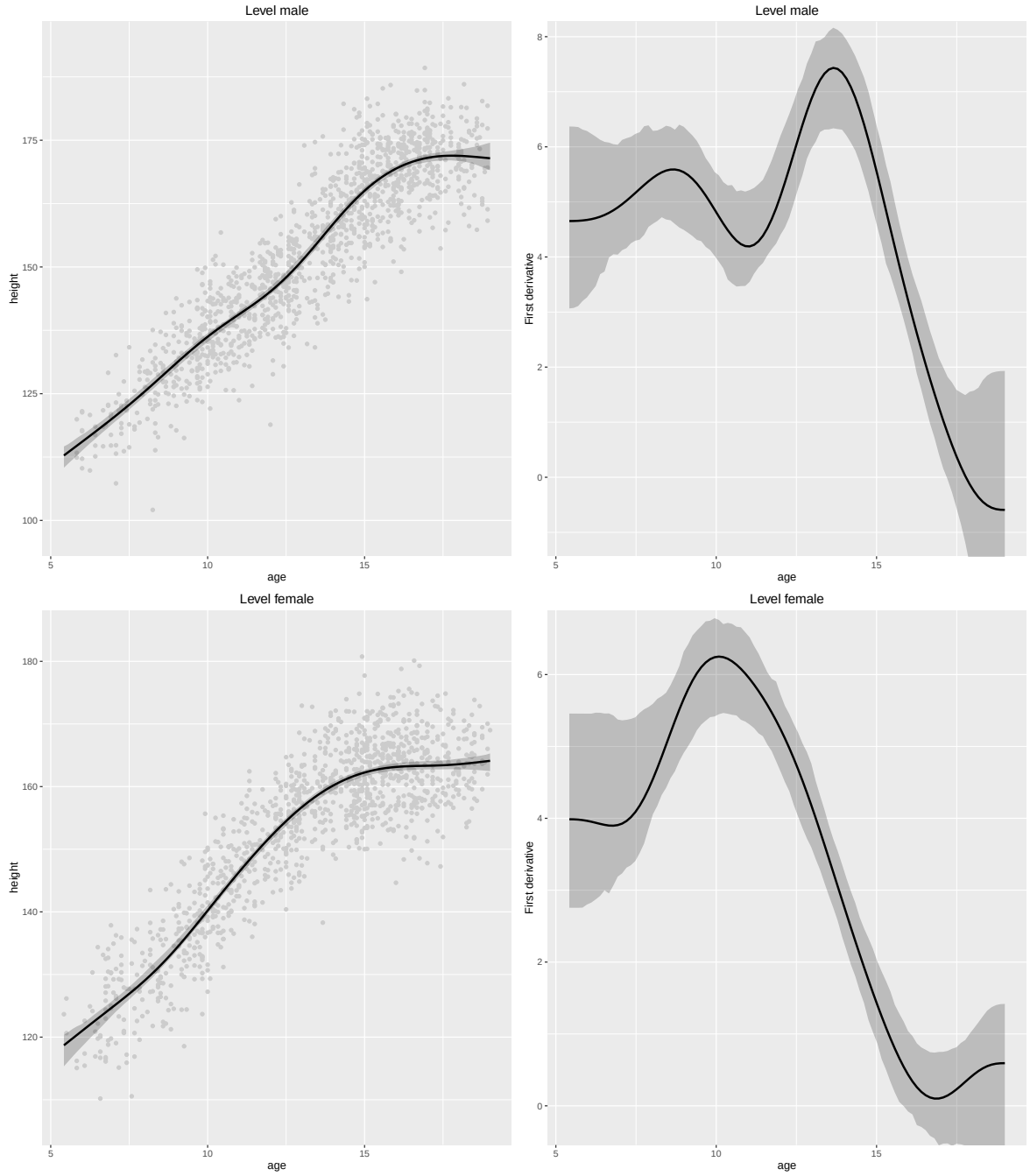

Figure 5: Regression curve and first derivative (solid lines) with bootstrap-based 95% pointwise confidence intervals (shaded area) for height and age of males (first row) and females (second row) using splines.

```
  Statistic pvalue Decision
1  515.6921      0 Rejected
```

```
R> globaltest(height ~ age:sex, data = children, p = 2, seed = 130853,
+    der = 1)
```

```
  Statistic pvalue Decision
1  143.6067      0 Rejected
```

```
R> globaltest(height ~ s(age, by = sex), data = children, seed = 130853,
+    der = 1, smooth = "splines")
```

```
  Statistic pvalue Decision
1  165.5214      0 Rejected
```

In addition, Figures 4 and 5 (right panels) seem to suggest that females experience a spurt in growth earlier than males do, with the two sexes achieving maximum rates of growth at ages close to 10 and 13 years, respectively. These ages are obtained by the `critical` function using the following command:

```
R> critical(mwi2k, der = 1)
```

```
             Critical      Lwr      Upr
Level male   13.64412 5.841401 13.88881
Level female 10.47682 5.841401 10.84384
```

```
R> critical(mwi2s, der = 1)
```

```
             Critical      Lwr      Upr
Level male   13.65030 12.89243 14.19899
Level female 10.06111  9.40000 11.11889
```

Paying close attention to the obtained confidence intervals using kernel smoothers, it is possible to observe that the lower limit of both intervals coincides with the smaller binning node. This occurs due to the high variability of the estimates for ages lower than eight years, resulting in a wide bootstrap confidence interval in this area. According to this, the critical point of several bootstrap replicates is estimated as the first point of the distribution. Two arguments of the `frfast` function (`rankl` and `ranku`) have been proposed in order to address this situation. By specifying them, the user can set a range in which the critical point will be searched.

```
R> mwi3k <- frfast(height ~ age:sex, data = children, p = 2, seed = 130853,
+    rankl = 8, ranku = 15)
R> critical(mwi3k, der = 1)
```

```
             Critical       Lwr      Upr
Level male   13.64412 12.971581 13.88881
Level female 10.47682  9.022302 10.85744
```

Finally, in order to ensure that these differences are really significant, we apply the `localtest` function. It tests whether the points that maximize the first derivatives of the curves are equal. Judging by these results, the sex-related differences in growth seem to be evident (using kernel or splines smoothers).

```
R> localtest(height ~ age:sex, data = children, p = 2, seed = 130853,
+    der = 1, rankl = 8, ranku = 15)

       d   Lwr    Upr Decision
1 3.1673 2.366 4.6626 Rejected

R> localtest(height ~ s(age, by = sex), data = children, seed = 130853,
+    der = 1, smooth = "splines")

        d     Lwr     Upr Decision
1 -3.9538 -4.7719 -2.6092 Rejected
```

5. Conclusion

This paper discussed the implementation of some methods developed for estimating regression models with or without factor-by-curve interactions in the R package **npregfast**. Among other things, the package also implements two bootstrap-based procedures designed to test different features of the estimated curves, particularly, to analyze whether the specific curves for each level are equal and to test the equality of the critical points estimated from the respective level-specific curves.

Finally, users may be interested in viewing a live interactive demo of the package in order to see part of its capabilities before installing it. This is possible at http://sestelo.shinyapps.io/npregfast.

Acknowledgments

This work was supported by research grant SFRH/BPD/93928/2013 of "Fundação para a Ciência e a Tecnologia" (FCT) and by FEDER Funds through "Programa Operacional Factores de Competitividade – COMPETE", by Portuguese Funds through FCT, within Project UID/MAT/00013/2013, by grant MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science and Innovation and by grant 10PXIB300068PR from the Galician Regional Authority (Xunta de Galicia). We thank the editorial team and reviewers for their constructive comments.

References

Bernard FR (1988). "Potential Fishery for the Gooseneck Barnacle *Pollicipes Polymerus* (Sowerby, 1833) in British Columbia." *Fisheries Research*, **6**(3), 287–298. doi:10.1016/0165-7836(88)90020-3.

Bidegain G, Guinda X, Sestelo M, Roca-Pardiñas J, Puente A, Juanes JA (2015). "Assessing the Suitability of the Minimum Capture Size and Protection Regimes in the Gooseneck Barnacle Shellfishery." *Ocean & Coastal Management*, **104**, 150–158. doi:10.1016/j.ocecoaman.2014.12.015.

Bidegain G, Sestelo M, Roca-Pardiñas J, Juanes JA (2013). "Estimating a New Suitable Catch Size for Two Clam Species: Implications for Shellfishery Management." *Ocean & Coastal Management*, **71**, 52–63. doi:10.1016/j.ocecoaman.2012.09.009.

Bowman A, Young S (1996). "Graphical Comparison of Nonparametric Curves." *Journal of the Royal Statistical Society C*, **45**(1), 83–98. doi:10.2307/2986225.

Bowman AW, Jones MC, Gijbels I (1998). "Testing Monotonicity of Regression." *Journal of Computational and Graphical Statistics*, **7**(4), 489–500. doi:10.1080/10618600.1998.10474790.

Cadarso-Suárez C, Roca-Pardiñas J, Molenberghs G, Faes C, Nácher V, Ojeda S, Acuña C (2006). "Flexible Modelling of Neuron Firing Rates Across Different Experimental Conditions: An Application to Neural Activity in the Prefrontal Cortex during a Discrimination Task." *Journal of the Royal Statistical Society C*, **55**(4), 431–447. doi:10.1111/j.1467-9876.2006.00545.x.

Cardoso AC, Yule AB (1995). "Aspects of the Reproductive Biology of *Pollicipes Pollicipes* (Cirripedia; Lepadomorpha) from the Southwest Coast of Portugal." *Netherlands Journal of Aquatic Ecology*, **29**(3–4), 391–396. doi:10.1007/bf02084238.

Chaudhuri P, Marron JS (1997). "SiZer for Exploration of Structures in Curves." *Journal of the American Statistical Association*, **94**(447), 807–823. doi:10.1080/01621459.1999.10474186.

Coull BA, Ruppert D, Wand MP (2001). "Simple Incorporation of Interactions into Additive Models." *Biometrics*, **57**(2), 539–545. doi:10.1111/j.0006-341x.2001.00539.x.

Cruz T (1993). "Growth of *Pollicipes Pollicipes* (Gmelin, 1790) (Cirripedia, Lepadomorpha) on the SW Coast of Portugal." *Crustaceana*, **65**(2), 151–158. doi:10.1163/156854093x00522.

Cruz T (2000). *Biologia e Ecologia Do Percebe, Pollicipes Pollicipes (Gmelin, 1790), No Litoral Sudoeste Português*. Ph.D. thesis, Universidad de Évora.

del Río AQ, Estévez-Pérez G (2012). "Nonparametric Kernel Distribution Function Estimation with **kerdiest**: An R Package for Bandwidth Choice and Applications." *Journal of Statistical Software*, **50**(8), 1–21. doi:10.18637/jss.v050.i08.

Delgado MA (1993). "Testing the Equality of Nonparametric Regression Curves." *Statistics & Probability Letters*, **17**(3), 199–204. doi:10.1016/0167-7152(93)90167-h.

Dette H, Neumeyer N (2001). "Nonparametric Analysis of Covariance." *The Annals of Statistics*, **29**(5), 1361–1400.

Efron B (1979). "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, **7**(1), 1–26.

Efron E, Tibshirani RJ (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Fan J, Gijbels I (1996). *Local Polynomial Modelling and Its Applications*. Number 66 in Monographs on Statistics and Applied Probability Series. Chapman and Hall.

Fan J, Marron JS (1994). "Fast Implementation of Nonparametric Curve Estimators." *Journal of Computational and Graphical Statistics*, **3**(1), 35–56. doi:10.1080/10618600.1994.10474629.

Gehrke W (1995). *Fortran 95 Language Guide*.

Goldberg H (1984). "Posibilidades De Cultivo De Percebe, *Pollicipes Cornucopia* Leach, En Sistemas Flotantes." *Informes Técnicos del Instituto Español de Oceanografía*, **11**, 1–13.

Golub GH, Heath M, Wahba G (1979). "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter." *Technometrics*, **21**(2), 215–223. doi:10.2307/1268518.

González-Manteiga W, Crujeiras RM (2013). "An Updated Review of Goodness-of-Fit Tests for Regression Models." *TEST*, **22**(3), 361–411. doi:10.1007/s11749-013-0327-5.

Hall P, Hart JD (1990). "Bootstrap Test for Difference between Means in Nonparametric Regression." *Journal of the American Statistical Association*, **85**(412), 1039–1049. doi:10.1080/01621459.1990.10474974.

Härdle W, Mammen E (1993). "Comparing Nonparametric Versus Parametric Regression Fits." *The Annals of Statistics*, **21**(4), 1926–1947.

Härdle W, Marron JS (1990). "Semiparametric Comparison of Regression Curves." *The Annals of Statistics*, **18**(1), 63–89.

Hastie T, Tibshirani R (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hayfield T, Racine JS (2008). "Nonparametric Econometrics: The **np** Package." *Journal of Statistical Software*, **27**(5), 1–32. doi:10.18637/jss.v027.i05.

Herrmann E, Maechler M (2016). ***lokern**: Kernel Regression Smoothing with Local or Global Plug-In Bandwidth*. R package version 1.1-8, URL https://CRAN.R-project.org/package=lokern.

Kauermann G, Opsomer JD (2003). "Local Likelihood Estimation in Generalized Additive Models." *Scandinavian Journal of Statistics*, **30**(2), 317–337. doi:10.1111/1467-9469.00333.

Kulasekera KB (1995). "Comparison of Regression Curves Using Quasi-Residuals." *Journal of the American Statistical Association*, **90**(431), 1085–1093. doi:10.1080/01621459.1995.10476611.

Lavergne P (2001). "An Equality Test Across Nonparametric Regressions." *Journal of Econometrics*, **103**(1–2), 307–344. doi:10.1016/s0304-4076(01)00046-x.

Liu RY (1988). "Bootstrap Procedures Under Some Non-I.I.D. Models." *The Annals of Statistics*, **16**(4), 1696–1708.

Mammen E (1993). "Bootstrap and Wild Bootstrap for High Dimensional Linear Models." *The Annals of Statistics*, **21**(1), 255–285.

Mazza A, Punzo A (2014). "**DBKGrad**: An R Package for Mortality Rates Graduation by Discrete Beta Kernel Techniques." *Journal of Statistical Software*, **57**(2), 1–18. doi:10.18637/jss.v057.c02.

Mazza A, Punzo A, McGuire B (2014). "**KernSmoothIRT**: An R Package for Kernel Smoothing in Item Response Theory." *Journal of Statistical Software*, **58**(6), 1–34. doi:10.18637/jss.v058.i06.

Molares J, Freire J (2003). "Development and Perspectives for Community-Based Management of the Goose Barnacle (*Pollicipes Pollicipes*) Fisheries in Galicia (NW Spain)." *Fisheries Research*, **65**(1–3), 485–492. doi:10.1016/j.fishres.2003.09.034.

Neumeyer N, Dette H (2003). "Nonparametric Comparison of Regression Curves: An Empirical Process Approach." *The Annals of Statistics*, **31**(3), 880–920.

Pardo-Fernández JC, Van Keilegom I, González-Manteiga W (2007). "Testing for the Equality of k Regression Curves." *Statistica Sinica*, **17**(3), 1115–1137.

Park C, Kang KH (2008). "SiZer Analysis for the Comparison of Regression Curves." *Computational Statistics & Data Analysis*, **52**(8), 3954–3970. doi:10.1016/j.csda.2008.01.006.

Racine JS, Hart J, Li Q (2006). "Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models." *Econometric Reviews*, **25**(4), 523–544. doi:10.1080/07474930600972590.

Racine JS, Hayfield T (2017). ***np**: Nonparametric Kernel Smoothing Methods for Mixed Data Types*. R package version 0.60-3, URL https://CRAN.R-project.org/package=np.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roca-Pardiñas J, Cadarso-Suárez C, Nácher V, Acuña C (2006). "Bootstrap-Based Methods for Testing Factor-By-Curve Interactions in Generalized Additive Models: Assessing Prefrontal Cortex Neural Activity Related to Decision-Making." *Statistics in Medicine*, **25**(14), 2483–2501. doi:10.1002/sim.2415.

Ruppert D, Sheather SJ, Wand MP (1995). "An Effective Bandwidth Selector for Local Least Squares Regression." *Journal of the American Statistical Association*, **90**(432), 1257–1270. doi:10.2307/2291516.

Ruppert D, Wand MP (1994). "Multivariate Locally Weighted Least Squares Regression." *The Annals of Statistics*, **22**(3), 1346–1370.

Sestelo M (2013). *Development and Computational Implementation of Estimation and Inference Methods in Flexible Regression Models. Applications in Biology, Engineering and Environment.* Ph.D. thesis, Department of Statistics and O.R., University of Vigo.

Sestelo M, Roca-Pardiñas J (2011). "A New Approach to Estimation of Length-Weight Relationship of *Pollicipes Pollicipes* (Gmelin, 1789) on the Atlantic Coast of Galicia (Northwest Spain): Some Aspects of Its Biology and Management." *Journal of Shellfish Research*, **30**(3), 939–948. doi:10.2983/035.030.0336.

Sestelo M, Villanueva NM, Roca-Pardiñas J (2017). ***npregfast**: Nonparametric Estimation of Regression Models with Factor-by-Curve Interactions*. R package version 1.5.1, URL https://CRAN.R-project.org/package=npregfast.

Sonderegger D (2012). ***SiZer**: Significant Zero Crossings*. R package version 0.1-4, URL https://CRAN.R-project.org/package=SiZer.

Sparre P, Venema SC (1997). "Introduction to Tropical Fish Stock Assessment. Part 1. Manual." *FAO Fisheries Technical Paper 306/1*, Food and Agriculture Organization of the United Nations. Rev. 2.

Srihera R, Stute W (2010). "Nonparametric Comparison of Regression Functions." *Journal of Multivariate Analysis*, **101**(9), 2039–2059. doi:10.1016/j.jmva.2010.05.001.

Wand M (2015). ***KernSmooth**: Functions for Kernel Smoothing for Wand & Jones (1995)*. R package version 2.23-15, URL https://CRAN.R-project.org/package=KernSmooth.

Wand MP, Jones MC (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. 2nd edition. Chapman & Hall/CRC.

Wu CFJ (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis." *The Annals of Statistics*, **14**(4), 1261–1295.

Young SG, Bowman AW (1995). "Non-Parametric Analysis of Covariance." *Biometrics*, **51**(3), 920–931. doi:10.2307/2532993.

Affiliation:

Marta Sestelo
Centre of Mathematics
University of Minho, Portugal
SiDOR Research Group and CINBIO
University of Vigo, Spain
E-mail: sestelo@uvigo.es
URL: http://sestelo.github.io/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
December 2017, Volume 82, Issue 12 *Submitted:* 2016-02-12
doi:10.18637/jss.v082.i12 *Accepted:* 2017-01-30