Ann Inst Stat Math (2015) 67:411–436
DOI 10.1007/s10463-014-0463-z

Testing regression models with selection-biased data

J. L. Ojeda · W. González-Manteiga · J. A. Cristóbal

Received: 11 January 2013 / Revised: 25 February 2014 / Published online: 13 May 2014

Abstract In this paper, we study integrated regression techniques to check the adequacy of a given model in the context of selection-biased observations. We introduce integrated regression in this setting, providing not only a suitable statistic for enabling a model checking test, but also a bootstrap distributional approximation to carry out the test. We also address the behaviour of the test under different alternatives showing that this behaviour is asymptotically the same for both selection-biased and non selection-biased data. The technique is illustrated with a simulation study and a data analysis based on a real situation that shows the performance of the method and how selection bias affect both estimation and inference.

Keywords Bootstrap · Goodness of fit · Integrated regression · Selection-biased data · Marked empirical process

1 Introduction

The way sample data are observed plays a major role in any subsequent statistical derivation or study driven by the data. As is pointed out in Cox (1969), where a

J. L. Ojeda (✉) · J. A. Cristóbal
Dept. de Métodos Estadísticos, Fac. de Ciencias, U. de Zaragoza, Pedro Cerbuna num. 12,
50009 Zaragoza, Spain
e-mail: jojeda@unizar.es

J. A. Cristóbal
e-mail: cristo@unizar.es

W. González-Manteiga
Dept. de Estadística e Investigación Operativa, Fac. de Matemáticas, U. de Santiago de Compostela,
15782 Santiago de Compostela, Spain
e-mail: wenceslao.gonzalez@usc.es

number of sampling problems arising in an industrial setting are discussed, the direct observation of the phenomena of interest is not always possible. In such circumstances, we have to face the problem of extracting information on the basis of the observable phenomena about which we are able to register information.

Among the causes of the lack of direct information from the phenomena of interest mentioned in Cox (1969) we find the absence of a framework where the sampling procedure takes place, the inaccessibility of part of the population of interest, or the complexity of the object to be sampled. In addition to these sampling problems, Cox (1969) also presented a number of inherent drawbacks we have to face in such a setting, one of these being the lack of correction and/or adjustment in most cases. These concerns are also shared by a number of other authors, see for example Patil and Rao (1978), Patil (1984), Quesenberry and Jewell (1986), Patil and Taillie (1989), Rao (1997), Cristóbal and Alcalá (2001).

As mentioned in all these references, a large number of situations where selection bias occurs can be addressed by means of weighted distributions because, in most cases, when the random phenomena of interest is distributed according to a random variable $\mathbf{X}$, possibly in $\mathbf{R}^d$, with c.d.f. F, the c.d.f. F^w of the observed random variable $\mathbf{X}^w$ is given by

$$\mathrm{d}F^w(\mathbf{x}) = \frac{w(\mathbf{x})\mathrm{d}F(\mathbf{x})}{\mu_w}, \tag{1}$$

where w is a known non-negative weight function that characterizes the selection bias and $\mu_w = \mathbf{E}\left[w(\mathbf{X})\right] = \int w(\mathbf{u})\,\mathrm{d}F(\mathbf{u}) > 0$. As a consequence, the influence of the way we observe the data on usual estimators depends on both the function w and the estimator being considered.

Against this background where data observation modifies the real frequency of events, we propose a procedure that enables us to perform model checking for the regression function in this context, allowing us to decide if a parametric model is suitable or not. The basis of this procedure is not the bias correction inherent to usual estimation procedures in this type of setup, but the compensation of the selection bias that is present in the observed sample. As can be seen from (1), the reciprocal of the function $w(\mathbf{x})$ can compensate the distortion caused by the selection bias in the original distribution F whenever $\mathbf{P}\{w(\mathbf{X}) > 0\} = 1$.

Model Checking for direct observations (i.i.d. samples from the r.v. $\mathbf{X}$) has been extensively studied from different perspectives in the literature. Hart (1997) offers an overview of some of the available techniques to perform goodness of fit test from the nonparametric point of view. Among the nonparametric techniques, it is worth mentioning the developments presented in Härdle and Mammen (1993), Stute (1997) Fan et al. (2001) and Van Keilegom et al. (2008). The idea in Härdle and Mammen (1993) is to use a *minimum distance* approach based on the comparison of the nonparametric and the parametric fits jointly with the aid of bootstrap resampling techniques. On the other hand, Fan et al. (2001) and Fan and Jiang (2007) introduced the "Generalized likelihood ratio" to test in the parametric and semi-parametric settings, proving that the so-called *Wilks phenomena* also holds for suitably chosen nonparametric alternatives. Van Keilegom et al. (2008) adopted a *distributional* point of view to the goodness of fit focusing on the *error distribution*. Their approach is based on the comparison of the properly scaled nonparametric and parametric error distributions by mean of Cramer–von Mises and Kolmogorov–Smirnov type statistics. Yet a completely different approach can be found in Stute (1997), where marked empirical process techniques are introduced to perform goodness of fit tests, see also Zhu (2005).

While these works and the references therein show how rich the goodness of fit literature is when data can be observed directly, this is not the case for selection-biased observations. Most of the work devoted to selection-biased data is focused on the unidimensional case and it is mostly based on the observed distribution, or on ad hoc procedures, see for example Rao (1997), Navarro et al. (2001) or Patil (2002) and the references therein. It was not until recently that the problem has been addressed from a broader perspective in Ojeda et al. (2008) or Ojeda and Keilegom (2009) in an effort to extend existing methods to this framework. While these approaches have proven to work in the framework of selection-biased data, they still rely on proper bandwidth selection procedures which can be avoided if the Marked Empirical Process works in the selection-biased setting.

In line with the aforementioned studies, and following ideas proposed in Cristóbal and Alcalá (2000) and Cristóbal et al. (2004), the main motivation for this work is to extend and investigate the performance of the integrated regression function introduced in Stute (1997) to carry out goodness of fit tests in a framework where data from the real phenomena of interest are not present. In particular, we will study the cumulated residuals process and the influence the interaction between the approximation error and the selection bias has in it. In order to achieve this we introduce the main definitions and estimators in the first section, leaving the development of the test and its bootstrap implementation for the second section. The third and fourth sections are devoted to a brief simulation study and data analysis example based on real data that allow us to show the performance of the method jointly with the effect and consequences selection bias may have from the point of view of estimation and inference.

2 Linear models for selection-biased data

2.1 Model, data and assumptions

Throughout the rest of the paper we will assume that the phenomena of interest $(\mathbf{X}, Y) \in \mathbf{R}^d \times \mathbf{R}$ is a continuous random vector distributed according to F, with joint density function f, so $\mathrm{d}F(\mathbf{x}, y) = f(\mathbf{x}, y)\,\mathrm{d}\mathbf{x}\,\mathrm{d}y$, and continuous regression function $m(\mathbf{x}) = \mathbf{E}[Y|\mathbf{X} = \mathbf{x}]$. In order to ease the presentation, to fit m we will consider the model

$$\mathcal{M} = \left\{ m(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} = \sum_{j=1}^{k} \beta_j g_j(\mathbf{x}) \ : \ \boldsymbol{\beta} \in \Omega \subset \mathbf{R}^k \right\},$$

where $\mathbf{g}(\mathbf{x})^{\mathrm{T}} = (g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))$, is a row vector of bounded continuous functions and $\boldsymbol{\beta}$ is the vector of linear combination coefficients $(\beta_1, \ldots, \beta_k) \in \Omega$, a compact subset in $\mathbf{R}^k$.

Provided that the functions g_j are suitable for representing m (i.e. $m \in \mathcal{M}$), we have to determine the value $\boldsymbol{\beta}_0$ such that $m(\mathbf{x}) = m\big(\mathbf{x}; \boldsymbol{\beta}_0\big)$. From this perspective, the problem that we will address is how to check that this class of functions is adequate to represent m. More precisely, we will consider the following hypothesis test:

$$H_0 : m \in \mathcal{M} \text{ vs. } H_1 : m \notin \mathcal{M}$$

when the available sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ does not come from the target population $(\mathbf{X}, Y)$ but from its selection-biased version $(\mathbf{X}^w, Y^w)$ whose density is given by

$$\mathrm{d}F^w(\mathbf{x}, y) = f^w(\mathbf{x}, y)\,\mathrm{d}\mathbf{x}\,\mathrm{d}y = \frac{w(\mathbf{x}, y)\, f(\mathbf{x}, y)}{\mu_w}\,\mathrm{d}\mathbf{x}\,\mathrm{d}y, \tag{2}$$

where $\mu_w = \mathbf{E}\,[w(\mathbf{X}, Y)] = \int w(\mathbf{u}, v) f(\mathbf{u}, v)\,\mathrm{d}\mathbf{u}\,\mathrm{d}v > 0$.

The fact that the observations in the sample are i.i.d. observations from $(\mathbf{X}^w, Y^w)$, the selection-biased version of the target population $(\mathbf{X}, Y)$, has a number of different consequences from which it is worth mentioning that

$$\mathbf{E}\left[Y^w|\mathbf{X}^w = \mathbf{x}\right] = m(\mathbf{x})\left(1 + \frac{\mathbf{Cov}\,[Y, w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}]}{m(\mathbf{x})\mathbf{E}\,[w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}]}\right),$$

which means that in this framework usual regression estimators are going to be biased because $\mathbf{Cov}\,[Y, w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}]$ is not null for arbitrary w functions. Furthermore, marginal densities for $\mathbf{X}$ and Y also change in the following way:

$$f^w_{\mathbf{X}}(\mathbf{x}) = \frac{\mathbf{E}\,[w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}]}{\mu_w} f_{\mathbf{X}}(\mathbf{x}), \quad f^w_Y(y) = \frac{\mathbf{E}\,[w(\mathbf{X}, Y)|Y = y]}{\mu_w} f_Y(y)$$

being $f_{\mathbf{X}}(\mathbf{x})$ and $f_Y(y)$ the marginal densities of $\mathbf{X}$ and Y, respectively.

Through the rest of the paper we will require following assumptions on the observed population $(\mathbf{X}^w, Y^w)$, w and $\mathcal{M}$:

A1 $\mathbf{P}\{w(\mathbf{X}, Y) > 0\} = 1$ and $\mathbf{E}\left[w(\mathbf{X}^w, Y^w)^{-2}\right] < +\infty$.
A2 The Matrix $\mathbf{L} = \mathbf{E}\left[\mathbf{g}(\mathbf{X})\mathbf{g}(\mathbf{X})^{\mathrm{T}}\right]$ is not singular.
B1 The function

$$v^w(\mathbf{x}) = \mathbf{E}\left[\left(\frac{Y^w - m(\mathbf{X}^w)}{w(\mathbf{X}^w, Y^w)}\right)^2\middle|\mathbf{X}^w = \mathbf{x}\right] \tag{3}$$

is integrable with respect to F^w.

Assumption A1 plays a crucial role to identify the whole distribution F when we observe F^w, for if $\mathbf{P}\{w(\mathbf{X}, Y) = 0\} > 0$ there will be part of the support of F whose probability mass would be inaccessible, see Gill et al. (1988) for further details. As a consequence of having completely identified F, any problem-related parametrization, correlation or independence issues related to the population covariables, etc. have to do with F and not with the observed distribution F^w. In particular, this means that if a parametrization is identifiable for F, it will also be identifiable when we have data from selection-biased distribution F^w and we use compensation to avoid the selection bias. At this point, it is important to notice that this is not the case in general, as any parameter in F will appear for example involved with μ_w in F_w. This is a very remarkable feature of compensation: it allows modeling of F when dealing with data from F^w.

2.2 Estimation and asymptotics under H_0

Following Cristóbal and Alcalá (2000) and Wu (2000), we use the reciprocal of $w_i = w(\mathbf{x}_i, y_i)$ as a weight in the least squares minimization problem:

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{w_i}\left(y_i - \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\beta}\right)^2, \tag{4}$$

where we can think of w_i as a "compensation" of the effect of selection bias in every observation $(\mathbf{x}_i, y_i)$. The solution for the estimation of the vector of coefficients is then

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{G})^{-1}\,\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{Y},$$

where $\mathbf{Y}$ is the column vector with observations, $\mathbf{G}$ is the $n \times k$ matrix with entries $g_j(\mathbf{x}_i)$ for $i = 1, \ldots, n$, $j = 1, \ldots, k$ and $\mathbf{B}$ is a diagonal matrix given by diag $(w_1^{-1}, \ldots, w_n^{-1})$.

Thus, if $\boldsymbol{\epsilon}$ denotes the column vector $(\epsilon_1, \ldots, \epsilon_n)$ with ϵ_i the *regression errors* $(y_i - m(\mathbf{x}_i))$, we have

Proposition 1 *If assumptions* A1 *and* A2 *are fulfilled and* $m \in \mathcal{M}$, *then the estimator* $\hat{\boldsymbol{\beta}}_n$ *admits the following almost sure expansion*:

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \mu_w \mathbf{L}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i}{w_i} + O_k\left(\frac{\log\log n}{n}\right). \tag{5}$$

As a consequence $\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + O_k\left(\sqrt{\log\log n/n}\right)$ *almost surely.*

Recall that previous result holds for basis of continuous functions like splines, etc. Furthermore, it also holds when we have discrete covariables with suitable notational changes.

The concept of *Integrated Regression*, i.e, $I(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} m(\mathbf{z})\,\mathrm{d}F_{\mathbf{X}}(\mathbf{z})$, introduced in Stute (1997), turns out to be the key for developing the test based on the residual accumulation. As a consequence of (2) and its implications for the marginals of $(\mathbf{X}^w, Y^w)$ we have that

$$I(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} m(\mathbf{z})\,\frac{\mu_w}{\mathbf{E}\left[w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{z}\right]}\,\mathrm{d}F_{\mathbf{X}}^{w}(\mathbf{z}), \tag{6}$$

where we have used the integral symbol with limits $-\infty$ and $\mathbf{x} = (x_1, \ldots, x_d)$ to denote an integral over the domain $(-\infty, x_1] \times \cdots \times (-\infty, x_d]$. $I(\mathbf{x})$ uniquely determines m in such a way that, in this setting were the observations are biased, there is no other way to compute $I(\mathbf{x})$.

Proposition 2 *The function* $h(\mathbf{x}) = \mu_w(\mathbf{E}\,[w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}])^{-1}m(\mathbf{x})$ *for* $\mathbf{x} \in \mathbf{R}^d$ *is the unique measurable function* F^w*-a.e. such that* $I(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} h(\mathbf{z})\,dF^w(\mathbf{z})$.

As in the unbiased case, its computation can be made simple with the use of the compensation:

$$I_n^w(\mathbf{x}) = \frac{1}{n}\overline{w}^H \sum_{i=1}^{n} \frac{1}{w_i} y_i \mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}},$$

where $\overline{w}^H = \left(n^{-1}\sum_{i=1}^{n} 1/w_i\right)^{-1}$ and $\mathbf{x} \le \mathbf{z}$ for vectors $\mathbf{x}$ and $\mathbf{z}$ with components $(x_1, \ldots, x_d)$ and $(z_1, \ldots, z_d)$ is understood as $\mathbf{x} \in (-\infty, z_1] \times \cdots \times (-\infty, z_d]$.

Proposition 3 *If assumption* A1 *is fulfilled*

$$\lim_{n\to\infty} I_n^w(\mathbf{x}) = I(\mathbf{x})$$

uniformly and almost surely.

Bearing in mind previous discussion about the Integrated Regression function, let us now focus on the cumulative residual process for the fit of model $\mathcal{M}$

$$R_n^w(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i}\hat{\epsilon}_i \mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i}\Big(y_i - m\Big(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_n\Big)\Big)\mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}}. \tag{7}$$

This process comprises the main features of the residuals $\hat{\epsilon}_i$ in the same way I_n^w comprises the main features of m; hence it is sound to study its behaviour to see if $\mathcal{M}$ is a suitable model for m.

R_n^w can be decomposed as $R_n^w(\mathbf{x}) = R_n^{w^0}(\mathbf{x}) + R_n^{w^1}(\mathbf{x})$ with

$$\begin{aligned} R_n^{w^0}(\mathbf{x}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i}(y_i - m(\mathbf{x}_i))\mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}}, \\ R_n^{w^1}(\mathbf{x}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i}\Big(m(\mathbf{x}_i) - m\Big(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_n\Big)\Big)\mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i}\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\Big(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n\Big)\mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}}, \end{aligned}$$

where last equality is a consequence of (5) when $m \in \mathcal{M}$.

From previous expressions we can see that R_n^w cumulates two sources of error. The first accounts for the random error the data has in itself, and it depends on the regression errors ϵ_i. The second, which depends on $\hat{\boldsymbol{\beta}}_n$, has to do with the estimation error for the regression function. As a consequence, the first of these two components is unavoidable and inherent to the random phenomena we are studying. On the other hand, the second component depends on how well the class of function $\mathcal{M}$ fits m.

Clearly, both $R_n^{w^0}$ and $R_n^{w^1}$ are dependent not only on ϵ_i, but also on the reciprocal of the w_i as these are introduced by the compensation technique.

Proposition 4 *If assumptions* A1, A2 *and* B1 *are fulfilled*

$$R_n^{w^0}(\mathbf{x}) \to R_\infty^{w^0}(\mathbf{x})$$

in distribution in the space $D[\mathbf{R}]^d$, *where* $R_\infty^w(\mathbf{x})$ *is a Gaussian process with null expectation and whose covariance function is given by*

$$\mathbf{Cov}\left[R_\infty^{w^0}(\mathbf{x}), R_\infty^{w^0}\big(\mathbf{x}'\big)\right] = \mathbf{E}\left[\mathbf{1}_{\{\mathbf{X}^w \le \mathbf{x}\wedge\mathbf{x}'\}} v^w\big(\mathbf{X}^w\big)\right].$$

Having characterized the stochastic behaviour of $R_n^{w^0}(\mathbf{x})$, we can use it to address the distributional behaviour of $R_n^w(\mathbf{x})$ using the following uniform representation:

Proposition 5 *If assumptions* A1, A2 *and* B1 *are fulfilled*

$$R_n^w(\mathbf{x}) = R_n^{w^0}(\mathbf{x}) - \mathbf{G}(\mathbf{x})^{\mathrm{T}}\,\mathbf{L}^{-1}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{1}{w_i}\epsilon_i + O\left(\frac{\log\log n}{\sqrt{n}}\right)$$

almost surely and uniformly for $\mathbf{x} \in \mathbf{R}^d$, *where* $\mathbf{G}(\mathbf{x}) = \mathbf{E}\left[\mathbf{g}(\mathbf{X})\mathbf{1}_{\{\mathbf{X}\le\mathbf{x}\}}\right]$.

The following result establishes the asymptotic distribution for $R_n^w(\mathbf{x})$.

Theorem 1 *If assumptions* A1, A2 *and* B1 *are fulfilled, then when* $m \in \mathcal{M}$

$$R_n^w(\mathbf{x}) \to R_\infty^w(\mathbf{x})$$

in distribution in the space $D[\mathbf{R}]^d$, $R_\infty^w(\mathbf{x})$ *is a Gaussian process with null expectation and whose covariance function is given by* $K^{v^w}\big(\mathbf{x},\mathbf{x}'\big)$, *where*

$$\begin{aligned} K^h\big(\mathbf{x},\mathbf{x}'\big) = {} & \mathbf{E}\left[\mathbf{1}_{\{\mathbf{X}^w\le\mathbf{x}\wedge\mathbf{x}'\}}h\big(\mathbf{X}^w\big)\right] \\ & - \mathbf{G}\big(\mathbf{x}'\big)^{\mathrm{T}}\,\mathbf{L}^{-1}\,\mathbf{E}\left[\mathbf{1}_{\{\mathbf{X}^w\le\mathbf{x}\}}h\big(\mathbf{X}^w\big)\mathbf{g}\big(\mathbf{X}^w\big)\right] \\ & - \mathbf{G}(\mathbf{x})^{\mathrm{T}}\,\mathbf{L}^{-1}\,\mathbf{E}\left[\mathbf{1}_{\{\mathbf{X}^w\le\mathbf{x}'\}}h\big(\mathbf{X}^w\big)\mathbf{g}\big(\mathbf{X}^w\big)\right] \\ & + \mathbf{G}\big(\mathbf{x}'\big)^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{E}\left[h\big(\mathbf{X}^w\big)\mathbf{g}\big(\mathbf{X}^w\big)\mathbf{g}\big(\mathbf{X}^w\big)^{\mathrm{T}}\right]\mathbf{L}^{-1}\mathbf{G}(\mathbf{x}). \end{aligned}$$

Even though it is out of the scope of this work, it is worth noticing that the result still holds when discrete covariates are present jointly with continuous covariates. Thus if we have a binary covariable F such that $p_k = \mathbf{P}\{F = k\}$ for $k = 1, 2$, we can consider our data being composed of two samples of sizes n_1 and n_2 such that $n = n_1 + n_2$, with their respective random processes $R_{1,n}^w$ and $R_{2,n}^w$ Theorem 1 ensures that they have limit processes $R_{1,\infty}^w$ and $R_{2,\infty}^w$, respectively. Hence, the process R_n^w, which can be written as $\sqrt{\hat{p}_1}\,R_{1,n}^w + \sqrt{\hat{p}_2}\,R_{2,n}^w$ for $\hat{p}_i = n_i/n$ converges to $\sqrt{p_1}\,R_{1,\infty}^w + \sqrt{p_2}\,R_{2,\infty}^w$ in the space $D[\mathbf{R}]^d$, being p_1 and p_2 the marginal probabilities of the discrete covariable.

The statistics we are going to consider to perform the test are

$$K_n = \sup_{\mathbf{x}\in\mathbf{R}^d} \left|R_n^w(\mathbf{x})\right|, \quad W_n^2 = \int_{\mathbf{R}^d} R_n^w(\mathbf{z})^2 \, \mathrm{d}F(\mathbf{z}).$$

As we can see, the stochastic behaviour of $R_n^w(\mathbf{x})$ does not allow for simple asymptotics for these statistics; therefore,, we will introduce a suitable bootstrap scheme to develop the test in the next section.

2.3 The test under the alternatives

As we have seen, the process R_n^w can be decomposed as the summation of the processes $R_n^{w^0}$ and $R_n^{w^1}$. Indeed, when $m \notin \mathcal{M}$, $R_n^{w^1}(\mathbf{x})$ can be further decomposed as $R_n^{w^{11}}(\mathbf{x}) + R_n^{w^{12}}(\mathbf{x})$ being

$$R_n^{w^{1,1}}(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i} \big(m(\mathbf{x}_i) - m\big(\mathbf{x}_i; \boldsymbol{\beta}'\big)\big) \mathbf{1}_{\{\mathbf{x}_i \leq \mathbf{x}\}},$$

$$R_n^{w^{1,2}}(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i} \Big(m\big(\mathbf{x}_i; \boldsymbol{\beta}'\big) - m\Big(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_n\Big)\Big) \mathbf{1}_{\{\mathbf{x}_i \leq \mathbf{x}\}}$$

for $\boldsymbol{\beta}'$ the value of $\boldsymbol{\beta}$ which causes $\mathbf{E}\,[w(\mathbf{X}^w, Y^w)^{-1}(Y^w - m(\mathbf{X}^w; \boldsymbol{\beta}))^2]$ to be minimum, i.e. $(m\mathbf{x}_i; \boldsymbol{\beta}')$ is the closest function in $\mathcal{M}$ to the regression function m in the mean square error sense.

While in our previous analysis, $R_n^{w^{11}}(\mathbf{x}) = 0$ for all $\mathbf{x}$ because $\boldsymbol{\beta}' = \mathbf{L}^{-1}\mathbf{E}[\mathbf{g}(\mathbf{X})Y] = \boldsymbol{\beta}_0$, when $m \notin \mathcal{M}$ we also have to consider the *approximation error* $\Delta(\mathbf{x}) = m(\mathbf{x}) - m(\mathbf{x}; \boldsymbol{\beta}')$. Furthermore, as a consequence of the so-called "Least Squares Normal Equations" for the linear model, it turns out that $\mathbf{E}\,[w(\mathbf{X}^w, Y^w)^{-1}\mathbf{g}(\mathbf{X}^w)^{\mathrm{T}}\Delta(\mathbf{X}^w)] = 0$. This fact not only leads to the consistency of $\hat{\boldsymbol{\beta}}_n$ computed from the minimization problem in (4), but also to an expansion like the one given in Proposition 1 in terms of both the regression and the approximation error. The following results require an additional assumption on the approximation errors:

C1 The function

$$\mathbf{E}\left[\left(\frac{\Delta(\mathbf{X}^w)}{w(\mathbf{X}^w, Y^w)}\right)^2 \middle| \mathbf{X}^w = \mathbf{x}\right]$$

is integrable with respect to F^w.

Proposition 6 *If assumptions* A1*,* A2 *and* C1 *are fulfilled and* $m \notin \mathcal{M}$*, then* $\hat{\boldsymbol{\beta}}_n$ *admits the following almost sure expansion*:

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}' + \mu_w \mathbf{L}^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i) \frac{\epsilon_i'}{w_i} + O_k\left(\frac{\log\log n}{n}\right)$$

for $\epsilon_i' = \epsilon_i + \Delta(\mathbf{x}_i)$.

Therefore, we can characterize the stochastic behaviour of $R_n^w(\mathbf{x})$ by means of a stochastic process whose covariance resembles in some sense the covariance structure of $R_\infty^w(\mathbf{x})$, plus an additional drift term that has to do with the *cumulated approximation error* $D(\mathbf{x}) = \mu_w^{-1}\mathbf{E}\left[\Delta(\mathbf{X})\mathbf{1}_{\{\mathbf{X}\leq\mathbf{x}\}}\right]$, which is a consequence of the fact that $R_n^{w^{11}}(\mathbf{x})$ cumulates the *compensated approximation error* along the values of $\mathbf{x}_i$ when $m \notin \mathcal{M}$.

Theorem 2 *If assumptions* A1, A2, B1 *and* C1 *are fulfilled, when* $m \notin \mathcal{M}$

$$R_n^w(\mathbf{x}) - \sqrt{n}D(\mathbf{x}) \to R_\infty^{w,\Delta}(\mathbf{x})$$

in distribution in the space $D[\mathbf{R}]^d$ *where* $R_\infty^{w,\Delta}(\mathbf{x})$ *is a Gaussian process with null expectation and covariance function* $K^{v^{w,\Delta}}(\mathbf{x}, \mathbf{x}')$:

$$v^{w,\Delta}(\mathbf{x}) = \mathbf{E}\left[\left(\frac{Y^w - m(\mathbf{X}^w) + \Delta(\mathbf{X}^w)}{w(\mathbf{X}^w, Y^w)}\right)^2 \middle| \mathbf{X}^w = \mathbf{x}\right]. \tag{8}$$

Hence, in spite of the bias present in the data, we can conclude that under the *compensation* strategy selection bias in data does not affect the main terms order of the alternatives these tests are able to detect, but to the constants in those terms and the test power. While this is a consequence of the first order or mean preservation properties of compensation, it is clear from the expression given for $v^{w,\Delta}(\mathbf{x})$ that the cumulative residual process is also affected by both the reciprocal of the weight function and the approximation error. Indeed, in the unbiased data case $v^{w,\Delta}(\mathbf{x})$ is simply $v^w(\mathbf{x}) + \mathbf{E}\left[\Delta(\mathbf{X})\right]^2$ making it clear how model misspecification may affect the test power, but in the selection-biased case the expression is rather involved and the effects of regression errors, model misspecification and selection bias interact in a complex way.

The previous result enables us to prove the consistency of the test procedure.

Corollary 1 *If assumptions* A1, A2, B1 *and* C1 *are fulfilled, when* $m \notin \mathcal{M}$:

$$\mathbf{P}\{K_n > c\} \longrightarrow 1, \quad \mathbf{P}\left\{W_n^2 > c\right\} \longrightarrow 1$$

for any $c > 0$. *Hence the tests are consistent.*

As the main aim of the present work is to extend the results given in Stute (1997) to the framework of selection-biased data, we have only considered linear models in order to ease the presentation of the main results. While Linear Models allow for a simple exposition and understanding of the issues raised by selection-biased data when making inferences for the regression function, the results we have introduced are valid in a broader estimation context. As we can see, the assumptions required by the main results ensure the existence and uniqueness of the solution for the least squares minimization problem. These estimator properties can be obtained in a number of ways, see for example Jennrich (1969), or Stute (1997). In Ojeda and Keilegom

(2009) it is proved that, under a number of assumptions on the class of functions $\mathcal{M}$ and the population $(\mathbf{X}, Y)$, the compensation strategy also works for non linear least squares estimators in the selection-bias framework and hence the present results can also be extended also to that setting.

3 Bootstrap calibration

As it has been shown, the process R_n^w exhibits a complex covariance structure that makes it difficult to use such asymptotics. To overcome this problem, in this section, we introduce a combination of a bootstrap scheme and a suitable bootstrap parameter estimation procedure. The basic idea is to use the wild bootstrap approach (see Liu 1988; Stute et al. 1998 or Härdle and Mammen 1993) jointly with compensation. This leads to compensated residuals that allow us to obtain the appropriate stochastic behaviour for this context.

The resampling procedure we are going to consider is given by

$$\mathbf{x}_i^* = \mathbf{x}_i; \; y_i^* = m\Big(\mathbf{x}_i^*; \hat{\boldsymbol{\beta}}_n\Big) + \epsilon_i^*; \; \epsilon_i^* = \hat{\epsilon}_i \, \gamma_i, \tag{9}$$

where $\gamma_i \; i = 1, \ldots, n$ is an i.i.d. sample of the Wild Bootstrap variable Γ, which is independent of $(\mathbf{X}^w, Y^w)$ and has null expectation with variance and third moment equal to 1. Notice also that residuals $\hat{\epsilon}_i$ can be written as $\epsilon_i + \Delta(\mathbf{x}_i) + \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}(\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}_n)$, being $\boldsymbol{\beta}' = \boldsymbol{\beta}_0$ and $\Delta(\mathbf{x})$ the null function in case $m \in \mathcal{M}$.

Bearing in mind that we are interested in characterizing the stochastic behaviour of a process related to residuals, it is important to realize that, apart from any possible estimation and/or approximation error due to the lack of fit for m in class $\mathcal{M}$, the *bootstrap errors* ϵ_i^* are selection-biased because of being also computed at $(\mathbf{x}_i, y_i)$. As a consequence of this inherited selection-bias, the bootstrap estimator $\hat{\boldsymbol{\beta}}_n^*$ for the vector $\boldsymbol{\beta}$ of parameters in class $\mathcal{M}$ based on the bootstrap sample should be obtained from

$$\hat{\boldsymbol{\beta}}_n^* = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \frac{1}{w_i}\Big(y_i^* - \mathbf{g}\big(\mathbf{x}_i^*\big)^{\mathrm{T}}\boldsymbol{\beta}\Big)^2. \tag{10}$$

In this way, the use of wild bootstrap resampling together with the compensated estimator to avoid the selection bias leads to *bootstrap residuals* $\hat{\epsilon}_i^* = y_i^* - m(\mathbf{x}_i^*; \hat{\boldsymbol{\beta}}_n^*)$ whose accumulation resembles in a proper way the behaviour of R_n^w. Notice that although the expected value of $\hat{\epsilon}_i$ is non null when $m \notin \mathcal{M}$, the expected value of Γ is null, and $\hat{\boldsymbol{\beta}}_n^*$ follows the same sort of expansion we found for $\hat{\boldsymbol{\beta}}_n$ in Proposition 1.

Proposition 7 *If assumptions* A1*,* A2 *and* C1 *are fulfilled then the estimator* $\hat{\boldsymbol{\beta}}_n^*$ *admits the following almost sure expansion*:

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n + \mu_w \mathbf{L}^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i^*}{w_i} + O_k\left(\frac{\log\log n}{n}\right).$$

As a consequence $\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n + O_k\left(\sqrt{\log\log n/n}\right)$ *almost surely.*

The bootstrap counterpart of expression (7) is given by

$$R_n^{w*}(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i} \hat{\epsilon}_i^* \mathbf{1}_{\{\mathbf{x}_i \leq \mathbf{x}\}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i} \left(y_i^* - m\left(\mathbf{x}_i^*; \hat{\boldsymbol{\beta}}_n^*\right)\right) \mathbf{1}_{\{\mathbf{x}_i^* \leq \mathbf{x}\}}.$$

As y_i^* is built using the parametric fit found with (10), the bootstrap regression function always belongs to $\mathcal{M}$. Indeed *bootstrap residuals* $\hat{\epsilon}_i^*$, can be written as $\epsilon_i^* + \mathbf{g}(\mathbf{x}_i^*)^{\mathrm{T}} (\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)$, which is analogous to the expansion for $\hat{\epsilon}_i$ when $\Delta(\mathbf{x}) = 0$, (i.e.: $m \in \mathcal{M}$). Therefore, these residuals resemble the stochastic behaviour of the regression error under H_0 jointly with the compensation of the selection bias by means of the reciprocal of w_i leads to a process $R_n^{w*}(\mathbf{x})$ whose stochastic behaviour is useful to perform in a consistent manner the test no matter whether $m \in \mathcal{M}$ or $m \notin \mathcal{M}$.

Theorem 3 *Under the assumptions made in Theorem* 2:

$$R_n^{w*}(\mathbf{x}) \to R_\infty^{w,\Delta}(\mathbf{x})$$

in distribution in the space $D[\mathbf{R}]^d$.

In particular, when $m \in \mathcal{M}: \; R_n^{w*}(\mathbf{x}) \to R_\infty^{w}(\mathbf{x})$.

The last assertion in Theorem 3 states that the distributional behaviour for $R_n^{w*}(\mathbf{x})$ agrees with the one we found for $R_n^w(\mathbf{x})$ asymptotically in $D[\mathbf{R}]^d$. While the consistency of this bootstrap procedure follows from this last assertion, it is worth noticing that when $m \notin \mathcal{M}$ the tests still detect the deviations from H_0 because of $v^{w,\Delta}(\mathbf{x})$ (see (8)) being finite. Therefore, the bootstrap allow us to obtain quantiles for our testing statistics:

$$K_n^* = \sup_{\mathbf{x} \in \mathbf{R}^d} \left| R_n^{w*}(\mathbf{x}) \right|, \quad W_n^{2^*} = \int_{\mathbf{R}^d} R_n^{w*}(\mathbf{z})^2 \, \mathrm{d}F_n(\mathbf{z}).$$

Using the wild bootstrap resampling mechanisms we have just described we can obtain B bootstrap observations $(K_n^*)_j$ and $(W_n^{2^*})_j$ for $j = 1, \ldots, B$. The null hypothesis should be rejected if the proportion of the bootstrap samples that are larger than K_n and W_n^2, respectively, is less than the desired error level α.

This selection-bias adapted bootstrap scheme we have presented is crucial for obtaining a good calibration of the critical rejection point for the tests as a consequence of the complexity exhibited by the covariance of the process R_n^w.

4 Empirical study

The simulations carried out in this section are focused on the analysis of the acceptance/rejection performance of the test introduced in Sects. 2 and 3 when the data suffer from *length bias*, i.e. $w(\mathbf{x}, y) = y$. Besides models already studied in Ojeda et al. (2008) and Ojeda and Keilegom (2009), we also consider one of the examples in Stute et al. (1998) with suitable modifications to obtain a positive response. In this way, this empirical study will also provide the reader with a comparative view of existing methods.

In the examples that follow we will consider the regression function $\mathbf{E}\left[Y|\mathbf{X}=\mathbf{x}\right]$ to be $m(\mathbf{x})+A\delta(\mathbf{x})$ for $m\in\mathcal{M}$, a class of linear combinations of functions, and $\delta\notin\mathcal{M}$. Therefore, when $A\neq 0$ this regression function does not belong to $\mathcal{M}$. The scenarios we have contemplated have sample sizes $n=50$, 100, 200, and A taking the values in $(-a,a)$ for $a>0$ allowing in this way the exploration of different degrees of deviation from the null hypotheses. Some of the models include a parameter σ, with values 0.1 or 0.5 that controls the variance. The number of bootstrap replications is $B=1{,}000$ in all these scenarios and the empirical p values (i.e.: proportion of rejections) have been computed for 1,000 simulations. To ease the presentation we only include the tables for the example 4.2 when A is -1, 0, or 1, while for the rest of the scenarios we have plotted the empirical power function depending on the value of A when $n=200$.

Example 1 Model 3 in Stute et al. (1998) with multiplicative errors. This is a multivariate regression model in which the population $(\mathbf{X},Y)$ stochastic behaviour is given by

$$\begin{gathered}Y=(2+5X_1-X_2+A\,\delta(X_1,X_2))\Big(1+\sigma\,\mathcal{U}\Big(-\sqrt{3},\sqrt{3}\Big)\Big),\\(X_1,X_2)\sim\mathcal{U}(0,1)\times\mathcal{U}(0,1),\quad\delta(x_1,x_2)=x_1\,x_2.\end{gathered}$$

Notice the use of the multiplicative error to obtain a positive response value, which is needed to address length-biased sampling (i.e.: $w(\mathbf{x},y)=y$). As a consequence, of this multiplicative error $\mathbf{Var}\left[Y|\mathbf{X}=\mathbf{x}\right]$ depends on the regression function $m(\mathbf{x})$.

Example 2 Models in Ojeda et al. (2008). In this case, the population (X,Y) stochastic behaviour is given by

$$\begin{gathered}Y=\Big(2X-X^2+A\delta(X)\Big)\Big(1+0.1\,\mathcal{U}\Big(-\sqrt{3},\sqrt{3}\Big)\Big),\\X\sim\mathcal{U}(0,1)\end{gathered}$$

for δ being

$$\begin{aligned}\delta_1(x)&=\frac{1}{4}\exp\Big(-100(x-1/2)^2\Big),\\\delta_2(x)&=2(x-1/16)(x-1/2)(x-15/16).\end{aligned}$$

As before A controls the degree of separation of the null hypotheses (i.e.: quadratic model), but in this case we consider two different departures from it. While δ_1 is a local peak function, δ_2 exhibits a smooth and global cubic deviation, see Ojeda et al. (2008).

Example 3 First Model in Ojeda and Keilegom (2009). In this case, the population (X, Y) stochastic behaviour is given by

$$Y = 8 - 1.25X + 2(X - 1.5)^2 + A\delta(X) + \sigma\,\mathcal{U}(-5, 5),$$
$$X \sim \mathcal{U}(0, 5), \quad \delta(x) = (x - 1.5)^3.$$

The main difference in this case is that the errors are additive while in the previous ones they are multiplicative.

A quick view of each of the examples considered separately, see Figs. 1–4 and Table 1, shows the consistency of the test procedure as expected from the theoretical results. Thus the percentage of rejections increases as the absolute value of A increases, and this is better appreciated as n increases. In each of the empirical studies we can also see that when σ has a large value the test loses power.

While it is very difficult to compare the performance of the test procedures among these scenarios, there are a few facts that are worth noticing. The error distribution considered in the simulations, the uniform distribution, is, in some sense, a worst scenario case as the errors are no longer concentrated at zero. While it seems that in the case of multiplicative errors the tests have less power than in the case of the additive ones, this can be misleading, as a consequence of having very different ranges of perturbations and regressions considered in the examples.

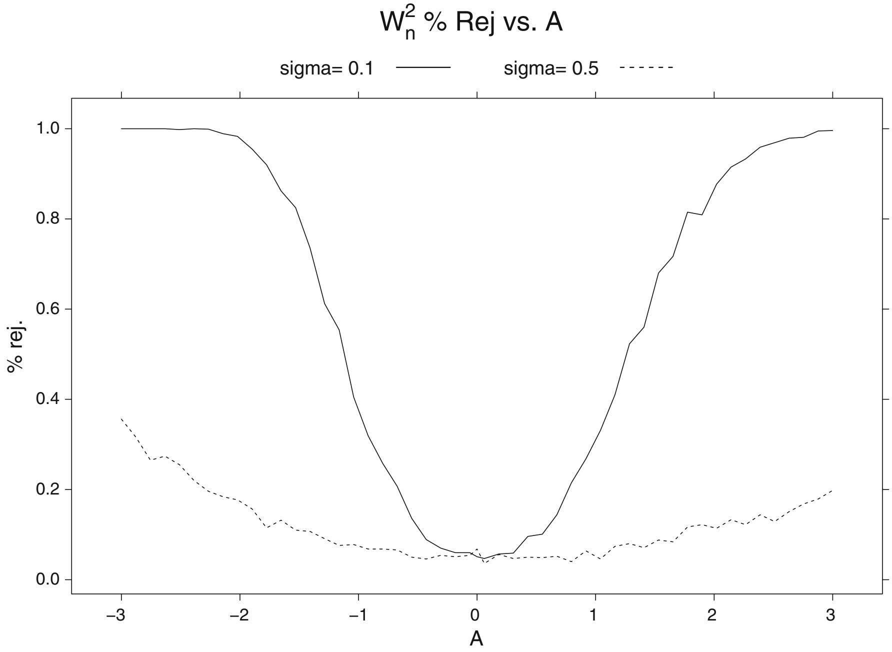

Fig. 1 Empirical power functions for W_n^2 in the example in Stute et al. (1998) ($\alpha = 0.05$)

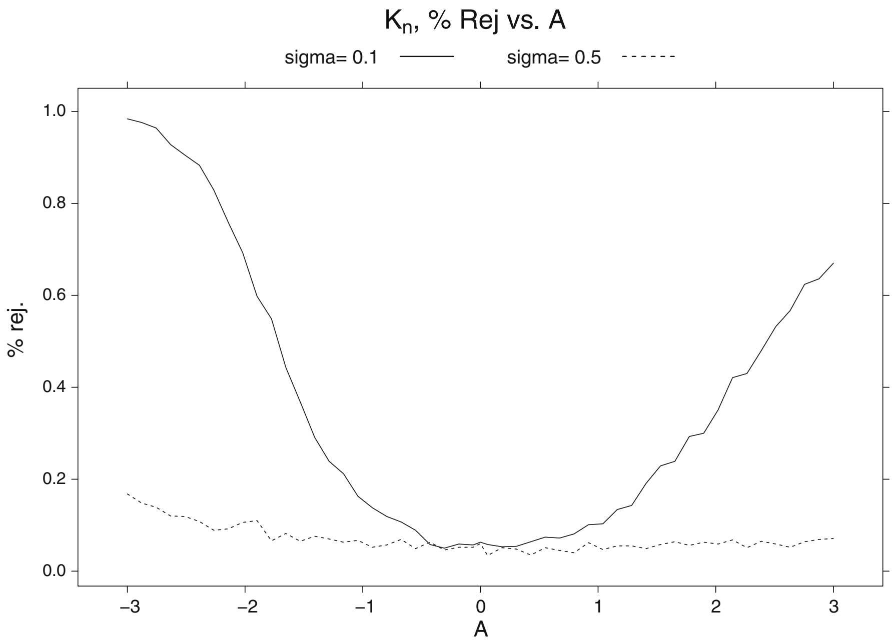

Fig. 2 Empirical power functions for K_n^∞ in the example in Stute et al. (1998) ($\alpha = 0.05$)

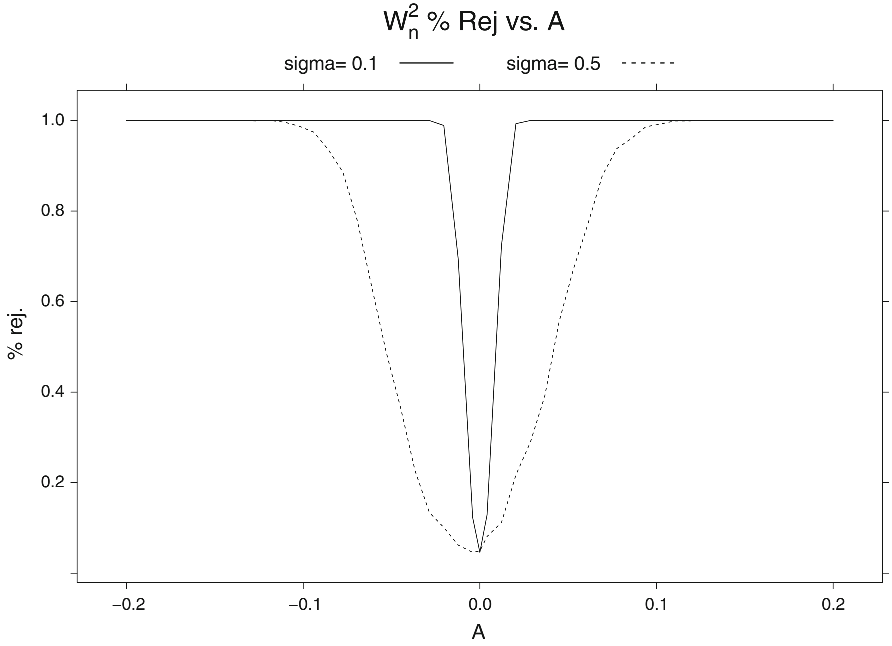

Fig. 3 Empirical power functions for W_n^2 in the example in Ojeda and Keilegom (2009) ($\alpha = 0.05$)

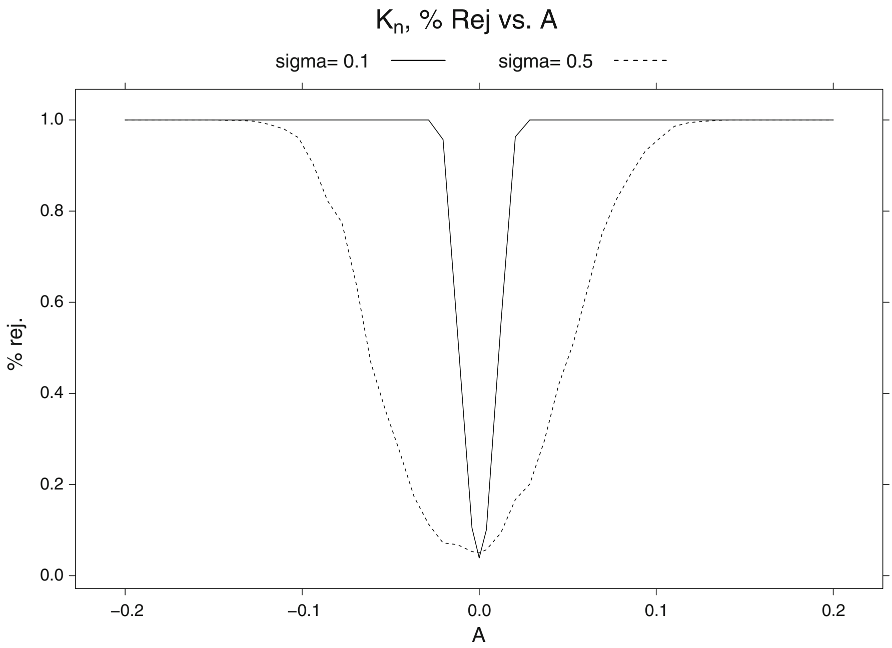

Fig. 4 Empirical power functions for K_n^{∞} in the example in Ojeda and Keilegom (2009) ($\alpha = 0.05$)

Example 4.2 deserves special attention as we have two different perturbations with the aim of characterizing both local and global departures from H_0. A direct comparison of the figures in Table 1 for $\alpha = 0.05$ with their counterparts in Ojeda et al. (2008) shows that as expected when $A = 0$ their behaviour is more or less similar for any n. On the other hand, when $A = 1$ it seems that the approach in Ojeda et al. (2008) gives a better performance for small samples.

In the case of the simulations carried out in Ojeda and Keilegom (2009), the comparisons between Figs. 3, 4 and the tables in that paper seem to suggest that the behaviour is quite similar.

5 Real case example

The aim of the following real case example is not only to show the empirical performance of the technique with real data, but also to emphasize the effects selection-bias has in estimation, inference and in the observed data. This real case example tackles *length-biased* sampling, a well known and studied selection bias, on a database of patients that require surgery. Notice, that without actual observations, i.e. with simulated data for example, it would not be possible to see how selection-biased sampling mechanism affects the whole process of sampling, estimation and inference. In particular, Figs. 5 and 6 below would have not been achievable.

Table 1 Rejection percentage of H_0 for K_n^∞ and W_n^2 depending on A for both perturbation functions δ_1 and δ_2 considered in the example in Ojeda et al. (2008)

δ	n	$\alpha = 0.01$			$\alpha = 0.05$		
		A	K_n^∞	W_n^2	A	K_n^∞	W_n^2
$\delta_1(x)$	50	−1.0	0.093	0.145	−1.0	0.303	0.405
		0.0	0.007	0.005	0.0	0.049	0.054
		1.0	0.114	0.169	1.0	0.350	0.419
	100	−1.0	0.349	0.473	−1.0	0.607	0.743
		0.0	0.012	0.010	0.0	0.045	0.054
		1.0	0.432	0.529	1.0	0.705	0.800
	200	−1.0	0.811	0.909	−1.0	0.937	0.975
		0.0	0.012	0.008	0.0	0.046	0.049
		1.0	0.855	0.934	1.0	0.958	0.983
$\delta_2(x)$	50	−1.0	0.091	0.178	−1.0	0.276	0.418
		0.0	0.006	0.004	0.0	0.052	0.036
		1.0	0.104	0.195	1.0	0.304	0.438
	100	−1.0	0.304	0.533	−1.0	0.598	0.783
		0.0	0.010	0.007	0.0	0.054	0.050
		1.0	0.354	0.588	1.0	0.643	0.831
	200	−1.0	0.752	0.931	−1.0	0.914	0.982
		0.0	0.011	0.010	0.0	0.043	0.039
		1.0	0.794	0.958	1.0	0.956	0.990

In order to achieve this, *length-bias* sampling is simulated in a database (population) consisting of male patients aged between 30 and 85 years for whom surgery was prescribed after *01-01-2001* and was carried out before *30-04-2001*. This database has been provided by the *Servicio Aragonés de Salud*, (the Aragon regional health authority), and it comprises information relating to the following variables:

- `FL`: Date of surgery prescription.
- `FS`: Date of surgical operation.
- `waiTim`: Time spent in the system. Number of days between the date surgery was prescribed and the date it was carried out.
- `age`: Age of the patient in years.

We will focus on the relationship between `waiTim` and `age`, modeling the dependence of the former on the latter.

Figure 5 provides a graphical representation of the variables `FL`, `FS` and `waiTim` for the entire population. Time is plotted in the x axis, so the time spent in the health system queue is represented by a straight line parallel to this axis. Individuals, who are sorted according to their surgical operation date, are placed the y axis. Usual unbiased sampling mechanism would select random individuals from the y axis, but when the sample consists on those patients that were in the system (i.e. patients waiting for the surgery to take place) on *1st. March, 2001* (vertical line located around the middle of the x axis) we obtain a length-biased sample regarding `waiTim`. The effect of

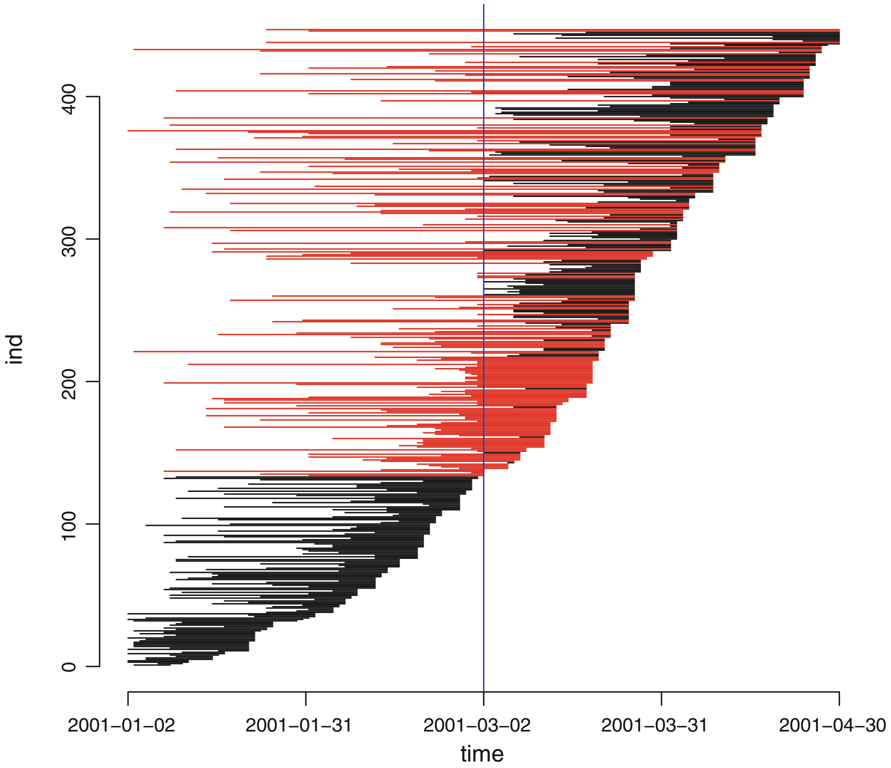

Fig. 5 Population with sampled individuals in *red* (colour figure online)

length bias is clear: those individuals whose duration are shown in red appear more frequently in the upper part of the plot, where those individuals with large duration are located.

Figure 6 is a scatter-plot of the variables `waiTim` and `age` for the "entire" database where sampled individuals have been coloured in red jointly with the *true* regression function and its length-biased simple linear regressions estimator. What we have called the *true* regression function is the Local Linear regression estimator over the entire database. This is plotted in black, jointly with its non parametric confidence bands, see Xia (1998), in green colour. The length-biased simple linear regressions estimator is plotted in blue and it is computed using the *length-biased* sample without the reciprocals of the responses to compensate the length bias; hence it is really an estimator of $\mathbf{E}\left[Y^w|X^w = x\right]$, i.e. a length-biased estimator of m.

As we have pointed out, the effect of *length bias* is visible in Fig. 5, where we can see that the larger `waiTim` is, the more chances an individual has to be in the *length-biased* sample. Nevertheless, the effect of the *length-biased* data on the regression estimators is much more important, we can see noticeable differences between the length-biased simple linear regression and the true population regression in Fig. 6. Indeed, the length-biased simple linear regressions estimator is outside the nonparametric confidence

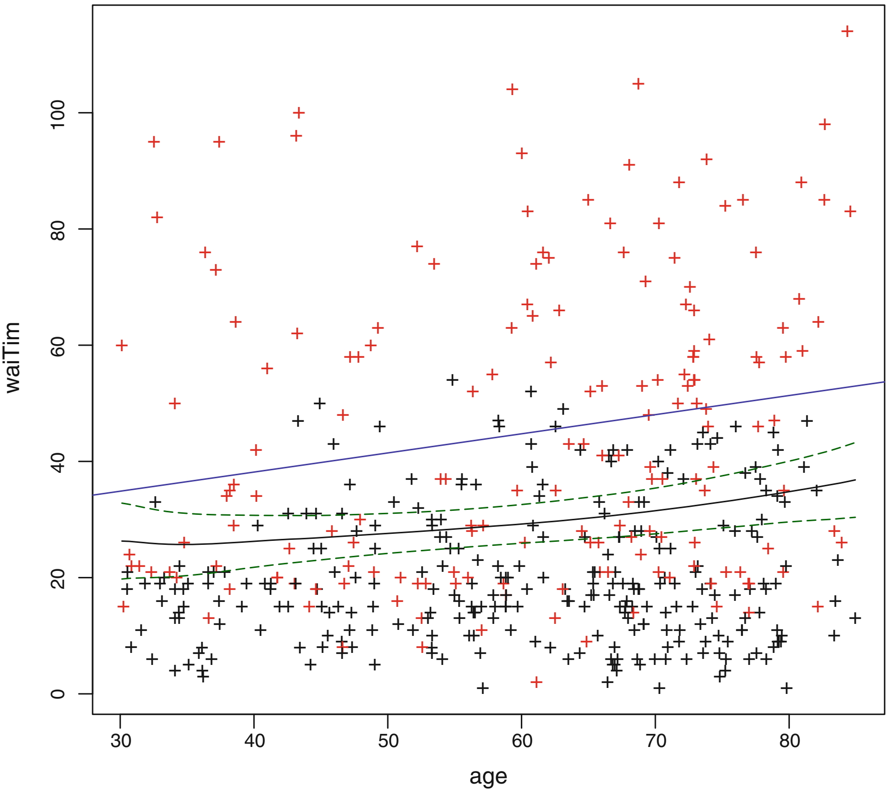

Fig. 6 The effect of length bias in estimation

Table 2 Int. Reg. compensated tests

	K_n p value	W_n^2 p value	K_n	W_n^2
`waiTim~1`	0.01	0.00	1.43	0.50
`waiTim~age`	0.27	0.19	0.80	0.17
`waiTim~-1+age`	0.17	0.11	1.14	0.36
`waiTim~-1+I(age^2)`	0.00	0.00	2.52	3.00
`waiTim~1+age+I(age^2)`	0.48	0.35	0.46	0.04

bands for the *true* regression function. It is worth noticing that these differences affect not only the intercept of the linear regression estimator, but also the slope.

The results for the different tests with statistics and p values are summarized in Table 2 for different Linear Models when the tests are developed using the reciprocal of the responses to compensate the *length bias* present in the data as described in Sects. 2.2 and 2.3, and the bootstrap procedure in Sect. 3 with $\text{B} = 999$ replications. When the significance level is $\alpha = 0.05$ all models with a linear term with variable `age` are accepted, even without intercept, which is consistent with what is shown in Fig. 6. In Table 3 we have included the results of the length-biased tests, that is to say the results of the tests without taking into account the *length bias* present in the

Table 3 Int. Reg. tests without compensation (length-biased tests)

	K_n p value	W_n^2 p value	K_n	W_n^2
`waiTim~1`	0.01	0.02	40.89	468.58
`waiTim~age`	0.47	0.53	15.22	37.47
`waiTim~-1+age`	0.02	0.01	45.62	1,016.65
`waiTim~-1+I(age^2)`	0.00	0.00	109.58	7,560.26
`waiTim~1+age+I(age^2)`	0.39	0.50	13.63	25.28

data. Notice that, apart from the differences existing in the values of the estimators, the linear model without intercept (model `waiTim`˜-1+age) is not accepted when the length bias is not considered. This means that, at the significance level $\alpha = 0.05$, the inference for the intercept would have been different if the *length bias* had not been taken into account.

Appendix

Proof of Proposition 1 (4) can be expressed in matrix terms as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{G}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{B}(\mathbf{Y} - \mathbf{G}\boldsymbol{\beta}),$$

therefore $\hat{\boldsymbol{\beta}}_n = (\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{Y}$. $n^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{G}$ is a matrix whose entries are $n^{-1}\sum_{i=1}^{n} w_i^{-1} g_j(\mathbf{x}_i) g_l(\mathbf{x}_i)$ and $\mathbf{E}\,[w(\mathbf{X}^w, Y^w)^{-1} g_j(\mathbf{X}^w) g_l(\mathbf{X}^w)] = \mu_w^{-1}\mathbf{E}\,[g_j(\mathbf{X}) g_l(\mathbf{X})]$ because of (1), plus all these entries have finite second-order moment as a consequence of assumption A1. Therefore, according to the Law of the Iterated Logarithm we have following expansion:

$$\frac{1}{n}\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{G} = \frac{1}{\mu_w}\mathbf{L} + O_{k\times k}\left(\sqrt{\frac{\log\log n}{n}}\right) \tag{11}$$

almost surely. As $\mathbf{Y} = \mathbf{G}^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$ for a sufficiently large n we obtain

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \left(\mathbf{G}^{\mathrm{T}}\mathbf{B}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{B}\boldsymbol{\epsilon}.$$

This, jointly with A2, leads to the following almost sure representation for $\hat{\boldsymbol{\beta}}_n$:

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \left(\mu_w\mathbf{L}^{-1} + O_k\left(\sqrt{\frac{\log\log n}{n}}\right)\right)\frac{1}{n}\mathbf{G}^{\mathrm{T}}\mathbf{B}\boldsymbol{\epsilon}.$$

Now, as $n^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{B}\boldsymbol{\epsilon}$ is a vector with entries $n^{-1}\sum_{i=1}^{n} w_i^{-1} g_j(\mathbf{x}_i)\epsilon_i$ the Law of the Iterated Logarithm means that $\mathbf{G}^{\mathrm{T}}\mathbf{B}\boldsymbol{\epsilon}$ is a matrix whose entries are quantities of order

$\sqrt{\log\log n/n}$ almost surely, leading to the almost sure representation given in the proposition. □

Proof of Proposition 2 If we assume that there exist another function h_1 such that $I(\mathbf{x}) = \int_{\infty}^{x} h_1(\mathbf{z})\,\mathrm{d}F^w_{\mathbf{X}}(\mathbf{z})$, then $\int_{\infty}^{\mathbf{x}}((h_1(\mathbf{z}) - h(\mathbf{z}))\mathrm{d}F^w_{\mathbf{X}}(\mathbf{z}) = 0$ for all $\mathbf{x}$ and $h_1(\mathbf{z}) = h(\mathbf{z})$ $F^w_{\mathbf{X}}$-a.e. □

Proof of Proposition 3 $I_n^w(\mathbf{x})$ can be written as an empirical process

$$I_n^w(\mathbf{x}) = \overline{w}^H \int \frac{1}{w(\mathbf{z}, y)}\mathbf{1}_{\{\mathbf{z}\le\mathbf{x}\}}\,\mathrm{d}F_n^w(\mathbf{z}, y) = \overline{w}^H\ F_n^w\ \frac{1}{w(\mathbf{z}, y)}\mathbf{1}_{\{\mathbf{z}\le\mathbf{x}\}}\,,$$

where $F_n^w\ f$ denotes the process $\int f(\mathbf{u}, v)\,\mathrm{d}F^w(\mathbf{u}, v)$ for f in a given class of functions. Therefore, $I_n^w(\mathbf{x})/\overline{w}^H$ is an empirical process indexed by the F^w-measurable VC-class of functions $\mathcal{C} = \{w(\mathbf{z}, y)^{-1}\mathbf{1}_{\{\mathbf{z}\le\mathbf{x}\}}\ :\ \mathbf{x} \in \mathbf{R}^d\}$ because they are indicators of quadrants in $\mathbf{R}^d$. As the envelope of $\mathcal{C}$ $e_{\mathcal{C}}(\mathbf{z}, y) = w(\mathbf{z}, y)^{-1}$ has finite expectation w.r.t. F^w, it verifies Glivenko–Cantelly property as stated in van der Vaart and Wellner (1996).

The Law of the Iterated Logarithm for the reciprocal of the responses $1/w_i$ proves that $\overline{w}^H - \mu_w$ is an $O\big(\sqrt{\log\log n/n}\big)$ quantity with probability one and the result follows. □

Proof of Proposition 4 From the definition of $R_n^{w^0}(\mathbf{x})$ it is clear that this process belongs to $D[\mathbf{R}]^d$ because any intersection between quadrants (see definition in Bickel and Wichura 1971) in $\mathbf{R}^d$ and $\{\mathbf{x} \in \mathbf{R}^d\ :\ \mathbf{x} \le \mathbf{x}_i\}$ is again a quadrant in $\mathbf{R}^d$ and $R_n^{w^0}(\mathbf{x})$ is a finite linear combination of the indicators of $\{\mathbf{x} \in \mathbf{R}^d\ :\ \mathbf{x} \le \mathbf{x}_i\}$ whose coefficients are continuous functions in $\mathbf{R}^d$. To prove the result we will follow Billingsley (1968) checking finite-dimensional distribution convergence and tightness.

The finite-dimensional distribution of a vector $(R_n^{w^0}(\mathbf{x}^1), \ldots, R_n^{w^0}(\mathbf{x}^k))$ for $\mathbf{x}^1, \ldots, \mathbf{x}^k$ in $\mathbf{R}^d$ is a multivariate normal distribution with null mean because for every $\mathbf{x}$ the expected value of $R_n^{w^0}(\mathbf{x})$ is null and covariance $\mathbf{E}[v^w(\mathbf{X}^w)\mathbf{1}_{\{\mathbf{X}^w\le\mathbf{x}\wedge\mathbf{x}'\}}]$ at $\mathbf{x}$ and $\mathbf{x}'$.

The proof of tightness will be based on the properties of the transformed process $Q_n^w(\mathbf{u})$ given by

$$R_n^{w^0}(\mathbf{x}) = Q_n^w(T(\mathbf{x})),$$

for

$$Q_n^w(\mathbf{u}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{w_i}\Big(y_i - m\Big(T^{-1}(\mathbf{u}_i)\Big)\Big)\mathbf{1}_{\{\mathbf{u}_i\le\mathbf{u}\}},$$

where $\mathbf{u}_i = T(\mathbf{x}_i)$, recall that as $w_i = w(\mathbf{x}_i, y_i)$ it also depends on $\mathbf{u}_i$. T defined as

$$T(\mathbf{x}) = \big(F^w(x_1|x_2, \ldots, x_d), F^w(x_2|x_3, \ldots, x_d), \ldots, F^w(x_{d-1}|x_d), F^w(x_d)\big),$$

where $F^w(x_i|x_{i+1}, \ldots, x_d)$ denotes the conditional distribution of the random variable $X_i^w|X_{i+1}^w, \ldots, X_d^w$ and $F^w(x_d)$ denotes the marginal distribution of X_d, the last variable in $\mathbf{X}$. As a consequence of this definition, T maps $\mathbf{R}^d$ into $[0, 1]^d$ and we will use F^q to denote the distribution of the transformed variable while $\mathbf{E}^q\,[\cdot]$ will be used for its expectation.

Bearing in mind the tightness criteria introduced in Theorem 3 in Bickel and Wichura (1971), the increment of a function H from $\mathbf{R}^d$ into $\mathbf{R}$ around a quadrant $D = [a_1, a_1 + b_1] \times \cdots \times [a_d, a_d + b_d]$ in $\mathbf{R}^d$ is defined as

$$H(D) = \sum_{l_1=0}^{1} \cdots \sum_{l_d=0}^{1} (-1)^{d-\sum_j l_j} H(a_1 + l_1 b_1, \ldots, a_d + l_d b_d).$$

Therefore, if $H(\mathbf{x}) = \mathbf{1}_{\{\mathbf{x}_j \leq \mathbf{x}\}}$ we have that $H(D) = \mathbf{1}_{\{\mathbf{x}_j \in D\}}$ and as $Q_n^w(\mathbf{u})$ is a linear combination of indicators $\mathbf{1}_{\{\mathbf{u}_i \leq \mathbf{u}\}}$, we obtain that for a quadrant $D \subset [0, 1]^d$

$$Q_n^w(D) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i} \Big(y_i - m\Big(T^{-1}(\mathbf{u}_i)\Big)\Big) \mathbf{1}_{\{\mathbf{u}_i \in D\}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{w_i} \alpha_i(D).$$

Hence, if quadrants D_1 and D_2 are neighbouring blocks in $[0, 1]^d$ (see definition in pg. 1658 in Bickel and Wichura 1971):

$$Q_n^w(D_1)^2 Q_n^w(D_2)^2 = \frac{1}{n^2} \left(\sum_{i=1}^{n} \alpha_i(D_1) \right)^2 \left(\sum_{i=1}^{n} \alpha_i(D_2) \right)^2.$$

Lemma 5.1 in Stute (1997) with $\alpha_i = \alpha_i(D_1)$ and $\beta_i = \alpha_i(D_2)$ leads to

$$\begin{aligned} \mathbf{E}^q \left[Q_n^w(D_1)^2 Q_n^w(D_2)^2 \right] \leq \frac{1}{n^2} & \Big(n\mathbf{E}^q \left[\alpha_i(D_1)^2 \alpha_i(D_2)^2 \right] \\ & + 3n(n-1)\mathbf{E}^q \left[\alpha_i(D_1)^2 \right] \mathbf{E}^q \left[\alpha_i(D_2)^2 \right] \Big). \end{aligned}$$

But as a consequence of D_1 and D_2 being disjoint sets we have that

$$\mathbf{E}^q \left[\alpha_i(D_1)^2 \alpha_i(D_2)^2 \right] = \mathbf{E}^q \left[\left(\frac{y_i - m\big(T^{-1}(\mathbf{u}_i)\big)}{w_i} \right)^2 \mathbf{1}_{\{\mathbf{u}_i \in D_1\}} \mathbf{1}_{\{\mathbf{u}_i \in D_2\}} \right] = 0$$

and, therefore,

$$\mathbf{E}^q \left[Q_n^w(D_1)^2 Q_n^w(D_2)^2 \right] \leq 3\frac{n-1}{n} \mathbf{E}^q \left[\alpha_i(D_1)^2 \right] \mathbf{E}^q \left[\alpha_i(D_2)^2 \right].$$

From which we have that

$$\mathbf{E}^q \left[\left| Q_n^w(D_1) \right|^2 \left| Q_n^w(D_2) \right|^2 \right] \leq \mu(D1)\, \mu(D_2),$$

where we have taken $\mu(D)$ to be $\sqrt{3}\mathbf{E}^q[\alpha_i(D)^2]$. Hence, condition (3) in Bickel and Wichura (1971) becomes fulfilled for neighbouring blocks D_1 and D_2, and Q_n^w is tight in $[0, 1]^d$ which means that R_n^w also is tight in $\mathbf{R}^d$.

Notice that μ is a measure that in this particular case is induced by the relative variance function v^w:

$$\mu(D) = \mathbf{E}^q\left[\left(\frac{y_i - m\big(T^{-1}(\mathbf{u}_i)\big)}{w_i}\right)^2 \mathbf{1}_{\{\mathbf{u}_i \in D\}}\right] = \int_{T^{-1}(D)} v^w(\mathbf{z})\,\mathrm{d}F_{\mathbf{X}}^w(\mathbf{z}).$$

□

Proof of Proposition 5 We will use the results given in Zhang (2006) to achieve a strong and uniform representation for

$$\mathbf{G}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_i}\mathbf{g}(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i \le \mathbf{x}\}} = \int \frac{1}{w(\mathbf{z}, y)}\mathbf{g}(\mathbf{z})\mathbf{1}_{\{\mathbf{z}\le\mathbf{x}\}}\mathrm{d}F_n^w(\mathbf{z}, y).$$

Notice that for every entry $g_j(\mathbf{x})$ in $\mathbf{g}(\mathbf{x})$, the integral can be written in terms of the empirical process theory as $F^w f$, for $f \in \mathcal{C}_j$ defined by

$$\mathcal{C}_j = \left\{ f(\mathbf{x}) \,:\, f(\mathbf{x}) = \frac{1}{w(\mathbf{z}, y)} g_j(\mathbf{z})\mathbf{1}_{\{\mathbf{z}\le\mathbf{x}\}},\ \mathbf{x} \in \mathbf{R}^d \right\}.$$

According to the notation in Zhang (2006), $P_{(n)}$(also $\overline{P}_{(n)}$) and P_n are given in this setting by F^w and F_n^w, respectively, while $\mathcal{F}$ is in our case $\mathcal{C}_j$. Both $\mu_{\mathcal{C}_j} = \| F^w f \|_{\mathcal{C}_j}$ and $\sigma^2_{\mathcal{C}_j} = \| F^w f^2 \|_{\mathcal{C}_j}$, where $\|\cdot\|_{\mathcal{C}_j}$ is the supremum over the class of functions $\mathcal{C}_j$, are finite as a consequence of g_j being uniformly bounded functions and assumption A1.

Using the arguments that were given in Proposition 3 for the class $\mathcal{C}$ with class $\mathcal{C}_j$ proves that $\mathcal{C}_j$ are VC-subgraph classes of functions, and hence covering numbers $N_2(\delta, F, \mathcal{C}_j)$ and $N_2(\delta, F_n, \mathcal{C}_j)$ are bounded by polynomials in δ. Furthermore, the functions in $\mathcal{C}_j$ are measurable in the product space determined by the sample of i.i.d. observations from $(\mathbf{X}^w, Y^w)$ and $\mathbf{R}^d$ that are indexed by $\mathbf{x} \in \mathbf{R}^d$, which is a complete metric space within $\mathbf{R}^d$ completion; therefore, they are permissible classes of functions (see Pollard 1984).

It remains to check $\| F_n^w f^2 - F^w f^2 \|_{\mathcal{C}_j} \longrightarrow 0$ to fulfill all assumptions required by Corollary 3.1 in Zhang (2006). Notice that for every $j = 1, \ldots, k$ the classes of functions

$$\mathcal{C}_j^* = \left\{ f(\mathbf{x}) \,:\, f(\mathbf{x}) = \frac{1}{w(\mathbf{z}, y)^2} g_j(\mathbf{z})^2\mathbf{1}_{\{\mathbf{z}\le\mathbf{x}\}},\ \mathbf{x} \in \mathbf{R}^d \right\}$$

are also VC-subgraph classes of F^w measurable functions with envelope $e_{\mathcal{C}}(\mathbf{z}, y) = w(\mathbf{z}, y)^{-2} g_j(\mathbf{z})^2$ arguing as in the proof of Proposition 3. As $e_{\mathcal{C}}(\mathbf{X}^w, Y^w)$ has finite expectation, $\mathcal{C}_j$ are Glivenko–Cantelly classes of functions and $\|F_n^w f^2 - F^w f^2\|_{\mathcal{C}_j} \longrightarrow 0$ almost surely. Therefore, Corollary 3.1 in Zhang (2006) implies that

$$\sup_{\mathbf{x}\in\mathbf{R}^d}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_i}g_j(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i\le\mathbf{x}\}} - \mathbf{E}\left[g_j(\mathbf{X})\mathbf{1}_{\{\mathbf{X}\le\mathbf{x}\}}\right]\right| = O\left(\sqrt{\frac{\log\log n}{n}}\right)$$

with probability one, and hence $\mathbf{G}_n(\mathbf{x}) = \mu_w^{-1}\mathbf{G}(\mathbf{x}) + O\left(\sqrt{\log\log n/n}\right)$ uniformly in $\mathbf{x}\in\mathbf{R}^d$ and almost surely.

Now the result follows from Proposition 1 since

$$\begin{aligned}R_n^w(\mathbf{x}) &= R_n^{w^0}(\mathbf{x}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{w_i}\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n\right)\mathbf{1}_{\{\mathbf{x}_i\le\mathbf{x}\}}\\ &= R_n^{w^0}(\mathbf{x}) - \frac{1}{\sqrt{n}}\mathbf{G}(\mathbf{x})^{\mathrm{T}}\mathbf{L}^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i}{w_i} + O\left(\frac{\log\log n}{\sqrt{n}}\right)\end{aligned}$$

almost surely and uniformly over $\mathbf{x}\in\mathbf{R}^d$. □

Proof of Theorem 1 As

$$R_n^w(\mathbf{x}) = R_n^{w^0}(\mathbf{x}) + R_n^{w^2}(\mathbf{x}) + O\left(\frac{\log\log n}{\sqrt{n}}\right)$$

almost surely and uniformly over $\mathbf{x}\in\mathbf{R}^d$ where

$$R_n^{w^2}(\mathbf{x}) = -\frac{1}{\sqrt{n}}\mathbf{G}(\mathbf{x})^{\mathrm{T}}\mathbf{L}^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i}{w_i},$$

and the stochastic behaviour of $R_n^{w^0}$ has been addressed in Proposition 4 we only need to study $R_n^{w^2}$.

The finite-dimensional distributions of $R_n^{w^2}(\mathbf{x})$ converge to finite-dimensional distributions of a Gaussian random process with null expectation and a covariance at $\mathbf{x}$ and $\mathbf{x}'$ given by $\mathbf{G}(\mathbf{x})^{\mathrm{T}}\mathbf{L}^{-1}\Sigma^w\mathbf{L}^{-1}\mathbf{G}(\mathbf{x}')$, with $\Sigma^w = \mathbf{E}\ [v^w(\mathbf{X}^w)\mathbf{g}(\mathbf{X}^w)\mathbf{g}(\mathbf{X}^w)^{\mathrm{T}}]$ because

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i}{w_i} \to N\left(\mathbf{0}, \Sigma^w\right)$$

as a consequence the multivariate CLT.

Tightness for $R_n^{w^2}(x)$ in $D[\mathbf{R}^d]$ can be proved as in Proposition 4 defining $Q_n^{w2}(\mathbf{u})$ by means of the same quantile transformation so that $R_n^{w^2}(\mathbf{x}) = Q_n^{w^2}(F^w(\mathbf{x}))$. The result follows having into account the covariance between processes $R_n^{w^0}(\mathbf{x})$ and $R_n^{w^2}(\mathbf{x})$. □

Proof of Proposition 6 The proof follows the same argumentation given in Proposition 1, but in this case, as $m \notin \mathcal{M}$ and $y_i = m(\mathbf{x}_i;\boldsymbol{\beta}')+m(\mathbf{x}_i)-m(\mathbf{x}_i;\boldsymbol{\beta}')+\epsilon_i$, we have to consider $\epsilon_i' = \epsilon_i + \Delta(\mathbf{x}_i)$. Recall that although ϵ_i' does not have null expectation, the expected value of $w_i^{-1}\mathbf{g}(\mathbf{x}_i)\Delta(\mathbf{x}_i)$ is still null because $w(\mathbf{X}^w, Y^w)^{-1}\mathbf{g}(\mathbf{X}^w)^{\mathrm{T}}\Delta(\mathbf{X}^w)$ has null expectation as a consequence of the way $\boldsymbol{\beta}'$ is defined and its variance is finite because of assumptions. □

Proof of Theorem 2 Notice that even though when $m \notin \mathcal{M}$ the process $R_n^w(\mathbf{x})$ has non null expectation Proposition 6 supports a strong and uniform representation like the one we have being studying for $R_n^w(\mathbf{x})$ when $m \in \mathcal{M}$ but using ϵ_i' instead of ϵ_i. Its centered version $\zeta(\mathbf{x}) = R_n^w(\mathbf{x}) - \sqrt{n}D(\mathbf{x})$ can be written as

$$\zeta_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\epsilon_i'}{w_i}\mathbf{1}_{\{\mathbf{x}_i\le\mathbf{x}\}} - D(\mathbf{x})\right) - \frac{1}{\sqrt{n}}\mathbf{G}(\mathbf{x})^{\mathrm{T}}\mathbf{L}^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i'}{w_i} + O\left(\frac{\log\log n}{\sqrt{n}}\right)$$

because $w(\mathbf{X}^w, Y^w)^{-1}\mathbf{g}(\mathbf{X}^w)^{\mathrm{T}}\Delta(\mathbf{X}^w)$ has null expectation and hence the first term in $R_n^w(\mathbf{x})$ is the only one that needs to be centered.

The rest of the proof proceeds as in Proposition 4 and Theorem 1 as the tightness and finite-dimensional weak convergence follows in the same way having into account that the conditional variance of $w_i^{-1}\epsilon_i'$ given $\mathbf{X}^w = \mathbf{x}_i$ is $v^{w,\Delta}(\mathbf{x}_i)$. □

Proof of Corollary 1 Recall that as $m \notin \mathcal{M}$, there exist $\mathbf{x}'$ such that $D(\mathbf{x}') = \mu_w^{-1}\int_{-\infty}^{\mathbf{x}'}\Delta(\mathbf{z})\,\mathrm{d}F(\mathbf{z}) \neq 0$. As $R_n^w(\mathbf{x}) = \zeta(\mathbf{x}) + \sqrt{n}D(\mathbf{x})$:

$$W_n^2 = \int_{\mathbf{R}^d}\zeta_n(\mathbf{x})^2\,\mathrm{d}F(\mathbf{z}) + n\int_{\mathbf{R}^d}D(\mathbf{z})^2\,\mathrm{d}F(\mathbf{z}) + o_p(n).$$

As $\int_{\mathbf{R}^d}D(\mathbf{z})^2\,\mathrm{d}F(\mathbf{z}) > 0$ because D is a continuous function and $D(\mathbf{x}) \neq 0$ at some neigbourhood V of $\mathbf{x}'$, we have that $\mathbf{P}\{W_n^2 > c\}\longrightarrow 1$ for any $c > 0$.

In the case of K_n, notice that when $m \notin \mathcal{M}$ we have $n^{-1/2}R_n^w(\mathbf{x}')\xrightarrow{p}\mu_w^{-1}D(\mathbf{x}')$, therefore for $c > 0$ we have

$$\mathbf{P}\{K_n > c\} \geq \mathbf{P}\left\{\left|R_n^w\left(\mathbf{x}'\right)\right| > c\right\}\longrightarrow 1.$$

□

Proof of Proposition 7 Reasoning as in Proposition 1 we have

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n + \mu_w\mathbf{L}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i^*}{w_i} + O_k\left(\frac{\log\log n}{n}\right).$$

When $m \in \mathcal{M}$, $\hat{\epsilon}_i = \epsilon_i + \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n)$ and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i^*}{w_i} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i}{w_i}\gamma_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\frac{\gamma_i}{w_i}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n\right).$$

The properties of γ_i, $\mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}$ jointly with the LIL lead to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\frac{\gamma_i}{w_i} = O_{k\times k}\left(\sqrt{\frac{\log\log n}{n}}\right),$$

and the result follows having into account that $\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n = O_k\left(\sqrt{\log\log n/n}\right)$ because of Proposition 1.

Recall that when $m \notin \mathcal{M}$, $\hat{\epsilon}_i = \epsilon_i' + \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}(\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}_n)$ and, as $\mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}'$ is the best approximation to $m(\mathbf{x})$ in $\mathcal{M}$ in the least squares sense and $\mathbf{E}[\mathbf{g}(\mathbf{X})^{\mathrm{T}}(\mathbf{Y} - \mathbf{g}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\beta}')] = 0$ and it has finite variance because of assumptions. Therefore, when $m \notin \mathcal{M}$, $\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}_n = O_k(\sqrt{\log\log n/n})$ and we can argue as in the case $m \in \mathcal{M}$. □

Proposition 8 *Under the assumptions made in Proposition 7 then*

$$R_n^{w*}(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\epsilon_i\gamma_i}{w_i}\mathbf{1}_{\{\mathbf{x}_i\leq\mathbf{x}\}} - \mathbf{G}(\mathbf{x})^{\mathrm{T}}\,\mathbf{L}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\frac{\epsilon_i\gamma_i}{w_i} + O\left(\frac{\log\log n}{\sqrt{n}}\right)$$

almost surely and uniformly for $\mathbf{x} \in \mathbf{R}$.

Proof The result follows as Proposition 5 expanding $\hat{\epsilon}_i^*$ in terms of ϵ_i and γ_i as $\gamma_i\epsilon_i + \gamma_i\Delta(\mathbf{x}_i) + \gamma_i\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}(\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}_n)$. In this case we also need to consider empirical processes indexed by following classes of functions:

$$\mathcal{C}_j' = \left\{f(\mathbf{x})\ :\ f(\mathbf{x}) = \frac{\gamma}{w(\mathbf{z}, y)}g_j(\mathbf{z})\mathbf{1}_{\{\mathbf{z}\leq\mathbf{x}\}},\ \mathbf{x} \in \mathbf{R}\right\}$$

with respect to the distribution of $(\mathbf{X}^w, \Gamma)$. As $\mathbf{X}^w$ and Γ are independent, its distribution is given by $\mathrm{d}F_{\mathbf{X}}^w(\mathbf{z})\ p_c$, being c equal to a or b, the values the wild bootstrap random variable Γ can take and p_a and p_b their respective probabilities. Proposition 6 ensures $\boldsymbol{\beta}' - \hat{\boldsymbol{\beta}}_n = O(n^{-1}\log\log n)$. □

Proof of Theorem 3 The result follows from Proposition 8 arguing as in Theorems 1 and 2, but in this case taking into account that we have to deal with $\epsilon_i'\gamma_i$ instead of ϵ_i.

Recall that $\epsilon_i\gamma_i$ follows a distribution determined by the r.v. $(\mathbf{X}^w, \varepsilon^w\Gamma)$ for $\varepsilon^w = (Y^w - m(X^w))$ which is a continuous r.v. whose distribution function is given by

$$\mathbf{P}\left(X^w \leq \mathbf{z}, \varepsilon^w\Gamma \leq e\right) = F^w\left(\mathbf{z}, \frac{e}{a}\right)p_a + F^w\left(\mathbf{z}, \frac{e}{b}\right)p_b.$$

Notice we have denoted by $F^w(\mathbf{z}, e)$ the distribution of the r. v. (X^w, ε^w). As a consequence, $\mathbf{E}\ [w_i^{-1}\epsilon_i'\gamma_i|\mathbf{X}^w = \mathbf{x}_i]$ is null and $\mathbf{E}\ [(w_i^{-1}\epsilon_i'\gamma_i)^2|\mathbf{X}^w = \mathbf{x}_i]$ is finite. □

Acknowledgments This work was partially supported by DGI Grants MTM2008-00625 and MTM2008-03010 (European FEDER support included), jointly with funds from the Consolidated Group DGA 2011-S11 and the DGA(CONSI + D) and the CAI EUROPA program. The authors would also like to thank the reviewers for their insightful comments and suggestions.

References

Bickel, P. J., Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Annals of Mathematical Statistics*, *42*, 1656–1670.

Billingsley, P. (1968). *Convergence of probability measures*. New York: John Wiley & Sons Inc.

Cox, D. R. (1969). Some sampling problems in technology. In N. L. Johnson & H. Smith (Eds.), *New Developments in Survey Sampling* (pp. 506–527). New York: Wiley.

Cristóbal, J. A., Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. *Journal of Statististical Planning and Inference*, *89*, 145–168.

Cristóbal, J. A., Alcalá, J. T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data. *Test*, *10*, 309–332.

Cristóbal, J. A., Ojeda, J. L., Alcalá, J. T. (2004). Confidence bands in nonparametric regression with length biased data. *Annals of the Institute of Statistical Mathematics*, *56*, 475–496.

Fan, J., Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *TEST*, *16*, 409–444.

Fan, J., Zhang, C., Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, *29*, 153–193.

Gill, R. D., Vardi, Y., Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, *16*, 1069–1112.

Härdle, W., Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, *21*, 1926–1947.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests*, Springer Series in Statistics. New York: Springer.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, *40*, 633–643.

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, *16*, 1696–1708.

Navarro, J., Ruiz, J. M., del Aguila, Y. (2001). Parametric estimation from weighted samples. *Biometrical Journal*, *43*, 297–311.

Ojeda, J. L., Cristóbal, J. A., Alcalá, J. T. (2008). A bootstrap approach to model checking for linear models under length-biased data. *Annals of the Institute of Statistical Mathematics*, *60*, 519–543.

Ojeda, J. L., Keilegom, I. V. (2009). Goodness-of-fit tests for parametric regression with selection biased data. *Journal of Statistical Planning and Inference*, *139*, 2836–2850.

Patil, G. (2002). Weigthed distributions. *Encyclopedia of Environmetrics*, *4*, 2369–2377.

Patil, G. P. (1984). Studies in statistical ecology involving weighted distributions. *Statistics: applications and new directions* (pp. 478–503). Calcutta: Indian Statist. Inst.

Patil, G. P., Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*, 179–189.

Patil, G. P., Taillie, C. (1989). Probing encountered data, meta analysis and weighted distribution methods. *Statistical data analysis and inference (Neuchâtel, 1989)* (pp. 317–345). Amsterdam: North-Holland.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Quesenberry, C. P, Jr, Jewell, N. P. (1986). Regression analysis based on stratified samples. *Biometrika*, *73*, 605–614.

Rao, C. R. (1997). *Statistics and truth*, 2nd edn. River Edge: World Scientific Publishing Co., Inc. (Putting chance to work, With a foreword by A. P. Mitra).

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, *25*, 613–641.

Stute, W., González Manteiga, W., Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, *93*, 141–149.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

Van Keilegom, I., González Manteiga, W., Sánchez Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test*, *17*, 401–415.

Wu, C. O. (2000). Local polynomial regression with selection biased data. *Statistica Sinica*, *10*, 789–817.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society Series B Statistical Methodology*, *60*, 797–811.

Zhang, D. X. (2006). Tail bounds for the supremums of empirical processes over unbounded classes of functions. *Acta Mathematica Sinica (English Series)*, *22*, 339–346.

Zhu, L. (2005). *Nonparametric Monte Carlo tests and their applications* (Vol. 182), Lecture Notes in Statistics. New York: Springer.