UNIVERSIDADE DE
SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE
ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

A bootstrap based model checking for selection-biased data

J. L. Ojeda, W. González-Manteiga, J. A. Cristobal

Report 07-05

Reports in Statistics and Operations Research

A BOOTSTRAP BASED MODEL CHECKING FOR SELECTION–BIASED DATA

J. L. Ojeda[1]
jojeda@unizar.es
W. González-Manteiga[2]
wences@zmat.usc.es
J. A. Cristóbal[1]
cristo@unizar.es

[1] Dept. de Métodos Estadísticos, U. de Zaragoza.

[2] Dept. de Estadística e I.O., U. de Santiago de Compostela

July, 2007

Abstract

In this paper, we study integrated regression techniques to check the adequacy of a given linear model in the context of selection biased observations. As a consequence, we introduce a definition of the integrated regression in this setting, giving not only a suitable statistic to perform a model checking test, but also a bootstrap distributional approximation to carry it out.

While the technique is introduced by means of the so–called length biased data, which is a particular case of selection bias, we also consider some other examples related to life–time analysis and reliability as well as stratification. All these examples are discussed in detail, addressing their main characteristics from the point of view of the selection bias they introduce. The paper ends with a brief simulation study that shows the empirical performance of the method.

Keywords: Bootstrap, Length-biased data, Goodness of fit, Integrated Regression, Marked Empirical Process.

1 Introduction

The way the sample data is 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 number of sampling problems that arise in the industrial setup are discussed, the direct observation of the phenomena of interest is not always possible. In those situations in which the straightforward observation of the phenomena is not accessible, we have to face the problem of extracting information on the basis of the observable phenomena we can register information about.

The sampling problems discussed in Cox (1969) were concerned with the way observations related to fibers, small–particles or work selection can be registered. What all these problems have in common is that often the way data is registered does not allow for the direct observation of the phenomena of interest, but the probability of recording an observation is proportional to its size. This complication, known as *length–bias* sampling, is a particular case of *selection–bias* sampling, where the frequency of the events we can observe does not agree with the frequency of those events of the real phenomena of interest. Among the causes for this drawback, Cox (1969) cited the absence of a framework where the sampling procedure takes place, the inaccessibility of part of the population of interest or simply the complexity of the object to be sampled. Besides all these causes for such disadvantages when sampling, he also mentions a number of inherent drawbacks we have to face in this setup, principally the lack of correction and/or adjustment in some cases. These concerns are also shared by a number of authors, see for example Quesenberry and Jewell (1986), Patil and Rao (1978), Patil (1984), Patil and Taillie (1989), Rao (1997), Cristóbal and Alcalá (2001).

As is 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 X with c.d.f. F, the observed random variable X^w c.d.f is given by

$$dF^w(x) = \frac{w(x)dF(x)}{\mu_w} \tag{1}$$

where $\mu_w = \mathbf{E}\left[w(X)\right] = \int w(x)\, dF(x)$. While this has a number of different consequences, let us stress the fact that, if $\mathbf{E}^w\left[\cdot\right]$ denotes the expectation with respect to the variable X^w, and t is a real function, then:

$$\mathbf{E}^w\left[t(X)\right] = \mathbf{E}\left[t(X)\right]\left(1 + \frac{\mathbf{Cov}\left[t(X), w(X)\right]}{\mathbf{E}\left[t(X)\right]\mathbf{E}\left[w(X)\right]}\right), \tag{2}$$

provided that the expectations with respect to X exist.

Therefore, we can conclude that the influence of the way we observe the data on the usual estimators depends on the covariance of the functions t and w, that is to say: the bias depends on the covariance between the function $t(X)$ we want information about, and $w(X)$ which depends on the way we observe the data. Notice that equation (2), while interesting in itself, does not allow for an immediate correction of the bias because we had to compute a covariance with respect to F that is not available.

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 for it. The basis of this procedure is not the bias correction that usual estimation procedures produce in this setup, but the compensation of the bias selection that is present in the observed sample. To be precise, whenever $w(x) > 0$, from (1) we have that the reciprocal of the function $w(x)$ at each of the observations from F^w can compensate the way the selection–bias distorts the original distribution of F. In this way we can recover the original distribution behavior making possible the estimation and inference about the regression function.

To be able to decide if a given parametric class of functions is a good choice to represent the regression function is one of the most interesting and challenging problems statistical theory faces. Model checking is not only crucial for making predictions but also for gaining true knowledge of the behavior of the phenomena we want information about. The model checking problem for direct observations (i.i.d. samples from the r.v. X), has been extensively studied from different perspectives in the literature, see among others, the book by Hart (1997), where a review of some of the available techniques of performing goodness of fit is offered, or Stute (1997), who introduces the Marked Empirical Process techniques for performing goodness of fit, see also Zhu (2005). It is also worth mentioning Härdle and Mammen (1993), where a comparison of the nonparametric and the parametric fits are used, and van Keilegom et al. (2007), who developed goodness of fit techniques based on the regression error estimation using nonparametric residuals. While the references included in all these works also show how rich the goodness of fit literature is in the usual framework, that is to say when data can be observed directly, this is not the case for selection biased observations. In the particular case of the selection biases we are going to deal with, see (1), most of the work we have found is devoted to the unidimensional case and uses the observed distribution as the basis for the developments, see for example Navarro et al. (2001), Rao (1997) or Patil (2002) and the references therein. As pointed out in Patil (2002), another interesting issue related to selection–bias is that in some circumstances and from the statistical point of view, it may be profitable to use biased observations, those from X^w, than to have direct observations from X, that is to say: we can also take advantage of biased observations in some circumstances.

In this work, and following ideas used in Cristóbal and Alcalá (2000), Cristóbal et al. (2004), and in Ojeda et al. (2004) to avoid the problems bias selected data introduces in the regression context we extend the work presented in Stute (1997) to this framework. Our main motivation is to study the utility and performance of the ideas about the integrated regression function in a framework where the data of the real phenomena under study is not present. In addition, we would like to highlight the fact that, as these ideas seem to work within quite different contexts, they offer a way of handling a number of different biased–selected data in an unified way.

To carry out this proposal, we will first deal with the bivariate response *length–biased* sampling case in Section 2. With this background, in Section 3 we will study some extensions starting from the case where more than a single covariate is present, and ending with some particular examples of *selection–bias* sampling. Section 4 will be devoted to a brief simulation study for the length–biased case. The proofs of the main results have been collected in a final Appendix.

2 Response Length Biased Data

Throughout the rest of this section we will assume that our population (X, Y) is a bivariate random variable whose distribution is F and whose joint density function is given by $dF(x, y) = f(x, y)dxdy$, in such a way that $Y > C > 0$ a.s. We will also assume that the regression function $m(x)$ admits a parametric linear representation, that is to say: it is a linear combination of given functions g_j:

$$m(x; \boldsymbol{\beta}) = \mathbf{g}(x)^T \boldsymbol{\beta} = \sum_{j=1}^{k} \beta_j g_j(x), \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of linear combination coefficients $(\beta_1, \dots, \beta_k) \in \Omega$, a compact in $\mathbf{R}^k$, and $\mathbf{g}(x)^T = (g_1(x), \dots, g_k(x))$, a row vector of functions. In this way, we can define a class of lineal models $\mathcal{M}$ as:

$$\mathcal{M} = \Big\{ m(x; \boldsymbol{\beta}) = \mathbf{g}(x)^T \boldsymbol{\beta} \, : \, \boldsymbol{\beta} \in \Omega \subset \mathbf{R}^k \Big\}, \tag{4}$$

and provided that the functions g_j are suitable to represent m (i.e. $m \in \mathcal{M}$), we have to determine the value $\boldsymbol{\beta}_0$ such that $\mathbf{E}\left[Y|X=x\right] = m(x) = m(x;\boldsymbol{\beta}_0)$.

In order to provide proper estimators for $\boldsymbol{\beta}_0$ we will assume the following additional hypothesis for the functions g_j:

A1 The Matrix

$$\mathbf{L} = \mathbf{E}\left[\mathbf{g}(X)\mathbf{g}(X)^T\right]$$

where $\mathbf{g}(X)$ is the column vector whose entries are $g_j(X)$, $j = 1,\ldots,k$ is not singular.

From the perspective of the class of linear functions $\mathcal{M}$, 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:

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

While we are interested in the regression function $m(x)$ that depends on the population distribution (X,Y), as a consequence of the length–bias sampling we cannot observe this random phenomena directly. Hence, instead of a random sample from the population we are interested in, the random sample $(x_1,y_1),\ldots,(x_n,y_n)$ we have comes from (X^{lb},Y^{lb}), the response length–biased version of (X,Y), whose density is given by:

$$dF^{lb}(x,y) = f^{lb}(x,y)\,dxdy = \frac{y\,f(x,y)}{\mu_Y}\,dxdy, \tag{5}$$

where $\mu_Y = \int yf(x,y)dxdy$. We will also denote the mean and variance for the observed data (i.e.: computed with distribution F^{lb}) by means of $\mathbf{E}^{lb}\left[\cdot\right]$ and $\mathbf{Var}^{lb}\left[\cdot\right]$ in order to distinguish them from $\mathbf{E}\left[\cdot\right]$ and $\mathbf{Var}\left[\cdot\right]$, which are those computed with distribution F. Therefore the probability of recording an observation (x,y) from the random phenomena (X,Y) we are interested in is proportional to y. Notice that, in this case

$$\mathbf{E}^{lb}\left[Y|X=x\right] = m(x)(1+c^2(x)),$$

and because of $c(x)$ being the conditional coefficient of variation, the direct application of the usual estimation techniques will lead to biased estimators. It is worth mentioning that the length–bias sampling also affects the marginals, and a simple computation leads to the following relationship between the observed (*length–biased*) marginal density for X^{lb} and the unobserved marginal density for X

$$f_X^{lb}(x) = \frac{m(x)f_X(x)}{\mu_Y}, \tag{6}$$

which tells us about the importance the regression function plays modifying the way covariates are observed under this kind of sampling.

As can be seen from equation (5), the reciprocal of the response variable can be used to compensate the length–bias. Therefore, and following Cristóbal and Alcalá (2000) and Wu (2000), we use the reciprocal of each observation as a weight in the usual Least Squares Minimization equation to obtain the following minimization problem:

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{y_i}\Big(y_i - \mathbf{g}(x_i)^T\boldsymbol{\beta}\Big)^2. \tag{7}$$

The solution for the estimation of the vector of coefficients is then:

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

where $\mathbf{Y}$ is the column vector with observations, $\mathbf{G}$ is the $n \times k$ matrix with entries $g_j(x_i)$ for $i = 1, \ldots, n$, $j = 1, \ldots, k$ and $\mathbf{B}$ is given by $\operatorname{diag}\big(y_1^{-1}, \ldots, y_n^{-1}\big)$. Under the assumptions made in the previous section, it can be proved that this estimator is strongly consistent. Thus, if $\boldsymbol{\epsilon}$ denotes the column vector $(\epsilon_1, \ldots, \epsilon_n)$ with ϵ_i the regression errors $(y_i - m(x_i))$, we have:

Proposition 2.1 *If assumption A1 is fulfilled and if the regression function m belongs to the class of functions $\mathcal{M}$, then the estimator $\hat{\boldsymbol{\beta}}_n$ admits the following almost sure expansion:*

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \mu_Y \mathbf{L}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i)\frac{\epsilon_i}{y_i} + O\left(\frac{\log\log n}{n}\right). \tag{8}$$

As a consequence,

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + O\left(\sqrt{\frac{\log\log n}{n}}\right)$$

almost surely and $\hat{\boldsymbol{\beta}}_n$ is a strongly consistent estimator for $\boldsymbol{\beta}_0$.

2.1 Integrated regression Estimation

In this framework, where sample observations are affected by length bias, the computation of the integrated regression is motivated by the relationship between the distributions F and F^{lb} that has just been explored. More precisely, notice that

$$I(x) = \int_{-\infty}^{x} m(z)\,dF(z) = \int_{-\infty}^{x} \mu_Y\,dF^{lb}(z) = \mu_Y F^{lb}(x). \tag{9}$$

because of $\mu_Y f_X^{lb}(x)$ being $m(x)f_X(x)$ as a consequence of equation (6). In fact, there is no other way to compute $I(x)$ in an integrated manner from F^{lb}.

Proposition 2.2 *The function $h(x) = \mu_Y$ for $x \in \mathbf{R}^d$ is the unique measurable function F^{lb}–a.e. such that*

$$I(x) = \int_{-\infty}^{x} h(z)\, dF^{lb}(z).$$

Hence,

$$I^{lb}(x) = \int_{-\infty}^{x} \mu_Y \, dF^{lb}(z),$$

uniquely determines $m(x)$, as happened in the unbiased case (see Stute (1997)) and, as $\mathbf{E}^{lb}\left[\mu_Y\, \mathbf{1}_{\{X\leq x\}}\right]$ is just $I^{lb}(x)$, its empirical counterpart is given by:

$$I_n^{lb}(x) = \frac{1}{n}\overline{y}^H \sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq x\}},$$

$\overline{y}^H$ being the harmonic mean, a strongly uniform estimator for μ_Y in this context. It is interesting to mention that this estimator can be viewed from the point of view of the compensation of the length–bias data exhibits by means of the reciprocal of the observations

$$I_n^{lb}(x) = \frac{1}{n}\overline{y}^H \sum_{i=1}^{n} \frac{1}{y_i}\, y_i\, \mathbf{1}_{\{x_i \leq x\}}.$$

On the basis of the latter expression, we can see that the role of the reciprocal of the observations seems to be a specialized case, or an adaptation to this setting of the so called *length–bias compensation* used in Cristóbal and Alcalá (2000) to perform local polynomial estimation in this context, and in Cristóbal et al. (2004) to build confidence bands for this kind of data.

We again find nice properties for $I_n^{lb}(x)$ in line with those given for $I_n(x)$ in Stute (1997).

Proposition 2.3 *Under the previous assumptions about (X, Y):*

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

uniformly and almost surely.

Bearing in mind the previous discussion about the way the estimation of $I(x)$ can be carried out in this framework, where the observations are length biased, the statistic we will use to perform goodness of fit will be based on the following marked empirical process

$$R_n^{lb^1}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{y_i} \Big(y_i - m\Big(x_i; \hat{\boldsymbol{\beta}}_n \Big) \Big) \mathbf{1}_{\{x_i \leq x\}}, \tag{10}$$

that can be written as

$$R_n^{lb^1}(x) = R_n^{lb}(x) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{y_i} \Big(m(x_i) - m\Big(x_i; \hat{\boldsymbol{\beta}}_n \Big) \Big) \mathbf{1}_{\{x_i \leq x\}},$$

where

$$R_n^{lb}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{y_i} (y_i - m(x_i)) \mathbf{1}_{\{x_i \leq x\}}.$$

We can see in this equation that $R_n^{lb^1}$ has two error sources. The first accounts for the random error the data has in itself, and it depends on the regression errors $\epsilon_i = y_i - m(x_i)$. The second has to do with the estimation error for the regression, that depends on the estimation error for $\boldsymbol{\beta}_0$.

Of course, the first of these two components is unavoidable and is inherent to the random phenomena we are studying. On the other hand, the second component depends basically on the class of function $\mathcal{M}$. In this way, if $m \notin \mathcal{M}$ we have that the differences between $m(x_i) - m\Big(x_i; \hat{\boldsymbol{\beta}}_n \Big)$ are accumulated, and hence $R_n^{lb^1}$ suffers a deviation from its behavior when $m \in \mathcal{M}$.

Besides all these considerations, it is interesting to point out that the estimation error depends on ϵ_i, and that the use of the compensation technique introduces the reciprocal of the responses in all these terms, which as we will see, has consequences on the variance of these estimators.

The stochastic behavior of $R_n^{lb}(x)$ can be characterized when $m \in \mathcal{M}$ provided that

B1 The random variable (X, Y) verifies

$$v^{lb}(x) = \mathbf{E}^{lb} \left[\left(\frac{Y - m(X)}{Y} \right)^2 \middle| X = x \right] \tag{11}$$

is integrable with respect to F^{lb}.

This requirement is a consequence of the compensation we use to obtain the estimators in this framework. Note that, as a consequence of B1, we have that

$$\mathbf{E}^{lb}\left[\left(\frac{Y-m(X)}{Y}\right)^2\right]<\infty.$$

Proposition 2.4 *If assumption B1 is fulfilled:*

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

in distribution in the space $D[-\infty,\infty]$, where $R_\infty^{lb}(x)$ is a gaussian process with null expectation and whose covariance function is given by

$$\mathbf{Cov}\left[R_\infty^{lb}(x),R_\infty^{lb}(x')\right]=\int_\infty^{x\wedge x'}v^{lb}(z)\,dF^{lb}(z).$$

Having characterized the stochastic behavior of $R_n^{lb}(x)$, we can use it to address the distributional behavior of $R_n^{lb^1}(x)$. In order to do this, recall that as a consequence of m belonging to the parametric class of functions $\mathcal{M}$ we can write:

$$R_n^{lb^1}(x)=R_n^{lb}(x)+\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{1}{y_i}\mathbf{g}(x_i)^T\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n\right)\mathbf{1}_{\{x_i\leq x\}},$$

and using expression (8), in the next Proposition we find a strong and uniform representation for $R_n^{lb^1}(x)$ that is more convenient for our purposes.

Let us define $\mathbf{G}(x)$ as:

$$\mathbf{G}(x)=\mathbf{E}\left[\mathbf{g}(X)\mathbf{1}_{\{X\leq x\}}\right]=\int_\infty^x\mathbf{g}(z)\,dF(z).$$

Proposition 2.5 *Under the assumptions made in Proposition 2.1 and assumption B1, if $g_1,\ldots,g_k$ are uniformly bounded functions in $\mathbf{R}$, then:*

$$R_n^{lb^1}(x)=R_n^{lb}(x)-\mathbf{G}(x)^T\,\mathbf{L}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{g}(x_i)\frac{1}{y_i}\epsilon_i+o(1)$$

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

This result enables us to obtain the main result of the section: the asymptotic distribution for $R_n^{lb^1}(x)$.

Theorem 2.1 *Under the assumptions made in Proposition 2.5:*

$$R_n^{lb^1}(x) \to R_\infty^{lb^1}(x)$$

in distribution in the space $D[-\infty,\infty]$, $R_\infty^{lb^1}(x)$ *is a gaussian process with null expectation and whose covariance function is given by*

$$\begin{aligned} K^{lb}\big(x,x'\big) &= \mathbf{Cov}\left[R_\infty^{lb^1}(x), R_\infty^{lb^1}\big(x'\big)\right] \\ &= \int_\infty^{x\wedge x'} v^{lb}(z)\, dF^{lb}(z) \\ &+ \int_\infty^{x} v^{lb}(z)\, \mathbf{G}\big(x'\big)^T\, \mathbf{L}^{-1}\mathbf{g}(z)\, dF^{lb}(z) \\ &+ \int_\infty^{x'} v^{lb}(z)\, \mathbf{G}(x)^T\, \mathbf{L}^{-1}\mathbf{g}(z)\, dF^{lb}(z) \\ &+ \mathbf{G}\big(x'\big)^T \mathbf{L}^{-1}\mathbf{\Sigma}^{lb}\mathbf{L}^{-1}\mathbf{G}(x), \end{aligned}$$

where

$$\mathbf{\Sigma}^{lb} = \mathbf{E}^{lb}\left[v^{lb}(X)\mathbf{g}(X)\mathbf{g}(X)^T\right].$$

2.2 Bootstrap Calibration

As we can see from the strong uniform approximation given in Proposition 2.5, the stochastic properties of $R_n^{lb^1}(x)$ are characterized by the functions $g_j(x)$, $j = 1,\ldots,k$ and the compensated residuals ϵ_i/y_i, $i = 1,\ldots,n$. In fact, the formula given in Theorem 2.1 for the covariance function of the weak limit of this process depends on $v^w(x)$ as a consequence of the way the compensation works to avoid the length–bias in the observations. Because of the complexity the covariance structure of this process exhibits, we will present a bootstrap scheme that, in some sense, follows the wild bootstrap approach (see Stute et al. (1998), Liu (1988) or Härdle and Mammen (1993)) with those compensated residuals we have mentioned above to get an appropriate stochastic behavior in this context where observations are length–biased.

The bootstrap sample we will use is given by:

$$x_i^* = x_i;\;\; y_i^* = m\Big(x_i^*;\hat{\boldsymbol\beta}_n\Big) + \hat\epsilon_i^*; \hat\epsilon_i^* = \hat\epsilon_i\,\gamma_i; \tag{12}$$

where $\hat\epsilon_i = y_i - \mathbf{g}(x_i)^T\hat{\boldsymbol\beta}_n$ and γ_i for $i = 1,\ldots,n$ is an i.i.d. sample of a random variable Γ, that is independent of the observed random sample (x_i, y_i), $i = 1,\ldots,n$, with null expectation, and variance and third moment equal to 1 (see Härdle and Mammen (1993) and the references therein). It is worth mentioning that in Stute et al. (1998)

some other bootstrap resampling methods were studied yielding that the classical bootstrap was inconsistent in the integrated regression setting. Their conclusion can also be extended to this setup.

Let us first focus on the consistency of the bootstrap estimator $\hat{\boldsymbol{\beta}}_n^*$ for the vector $\boldsymbol{\beta}$ of parameters in the class $\mathcal{M}$.

Proposition 2.6 *If assumption A1 is fulfilled then the estimator $\hat{\boldsymbol{\beta}}_n^*$ admits the following almost sure expansion:*

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n + \mu_Y \mathbf{L}^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{g}(x_i)\frac{\epsilon_i}{y_i}\gamma_i + O\left(\frac{\log\log n}{n}\right). \tag{13}$$

As a consequence,

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n + O\left(\sqrt{\frac{\log\log n}{n}}\right)$$

almost surely.

Following the expression given for $R_n^{lb^1}(x)$ in equation (10) we have that its bootstrap counterpart is given by

$$R_n^{lb^1*}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{y_i}\left(y_i^* - m\left(x_i^*; \hat{\boldsymbol{\beta}}_n^*\right)\right)\mathbf{1}_{\{x_i^* \leq x\}}. \tag{14}$$

Now as a consequence of $x_i^* = x_i$ we can write $R_n^{lb^1*}(x)$ in the following way:

$$R_n^{lb^1*}(x) = R_n^{lb*}(x) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{y_i}\mathbf{g}(x_i)^T\left(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*\right)\mathbf{1}_{\{x_i \leq x\}}.$$

where in this case

$$R_n^{lb*}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\hat{\epsilon}_i^*}{y_i}\mathbf{1}_{\{x_i \leq x\}}.$$

In this way, using residuals that completely resemble the stochastic behavior of the regression error, we are obtaining a stochastic behavior that is consistent in both situations, when the null hypothesis $H_0 : m \in \mathcal{M}$ is true, and also when the alternative hypothesis is true.

It is worth noticing that, as can be seen in all these expressions, in order to compensate the effect length–bias introduces in the bootstrap estimator, the true y_i values of the observed responses, not resampled responses y_i^*, should be used. This is a consequence of the fact that residuals $\hat{\epsilon}_i$ are biased, because of being computed using observations (x_i, y_i), that are distributed according to F^w. Finally, and because $x_i^* = x_i$ we have the following strong uniform representation for $R_n^{lb^1*}(x)$.

Proposition 2.7 *Under the assumptions made in Proposition 2.5 then:*

$$\begin{aligned} R_n^{lb^1 *}(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\epsilon_i}{y_i} \gamma_i \mathbf{1}_{\{x_i \le x\}} \\ &- \mathbf{G}(x)^T \mathbf{L}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{g}(x_i) \frac{\epsilon_i}{y_i} \gamma_i + o(1) \end{aligned}$$

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

Now the consistency of the bootstrap follows from the fact that the asymptotic distribution for $R_n^{lb*}(x)$ agrees with the one we found for $R_n^{lb}(x)$ in $D[-\infty, \infty]$.

Theorem 2.2 *Under the assumptions made in Theorem 2.1:*

$$R_n^{lb^1 *}(x) \to R_\infty^{lb^1}(x)$$

in distribution in the space $D[-\infty, \infty]$ where $R_\infty^{lb^1}(x)$ is a gaussian process with null expectation and covariance function $K^{lb}(x, x')$ as in Theorem 2.1.

As a consequence, the bootstrap integrated regression error mimics the stochastic behavior of the integrated regression error and we can use it to obtain quantiles for our testing statistics. The Bootstrap version of the statistics we are to use to perform the goodness of fit test are then given by

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

Now, using the wild bootstrap resampling mechanisms we have just described we can obtain B bootstrap observations $(K_n^*)_j$ and $\left(W_n^2\right)_j$ for $j = 1, \ldots, B$. The null hypothesis should be rejected if the proportion of the bootstrap samples that are larger than

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

respectively is less than the desired error level α. Therefore, it turns out that the bootstrap scheme we have just presented is crucial for obtaining a good calibration of the critical rejection point for our tests.

3 Some extensions

In this section we will present some extensions of previous work. First, we deal with covariates in $\mathbf{R}^d$, then we address the problem in the selection–bias case in a more general framework. The extensions will be presented in this way because of the ease of presentation: having understood the bivariate case it is simpler to understand the multivariate case, and then move to more general classes of selection–bias.

From now on we will assume that the population $(\mathbf{X}, Y)$ of interest is a multivariate random variable $(X_1, \ldots, X_d, Y)$ whose distribution is denoted by F and whose joint density function is given by

$$dF(\mathbf{x}, y) = f(\mathbf{x}, y)\, d\mathbf{x}dy = f(\mathbf{x}, y)\, dx_1 \ldots dx_d dy$$

and verifying that $Y > C > 0$ a.s. The regression function $m(\mathbf{x})$ is assumed to be a linear combination of given functions g_j from $\mathbf{R}^d$ on $\mathbf{R}$. Again the objective is to check if the functions g_j are suitable for representing the regression function m, and the statistics we will use are the extensions to the multivariate case of that proposed in the Section 2, the main change being that in this case the processes and the distribution F^{lb} are defined in the whole $\mathbf{R}^{d+1}$.

3.1 Response Length–bias with Multidimensional Covariates

As a consequence of the length–bias in the response, in this case we have

$$dF^{lb}(\mathbf{x}, y) = f^{lb}(\mathbf{x}, y)\, d\mathbf{x}dy = \frac{y\, f(\mathbf{x}, y)}{\mu_Y}\, d\mathbf{x}dy$$

As is natural, the assumptions A1 and B1 become in this case:

A1 The Matrix
$$\mathbf{L} = \mathbf{E}\left[\mathbf{g}(\mathbf{X})\mathbf{g}(\mathbf{X})^T\right]$$
where $\mathbf{g}(\mathbf{X})$ is the column vector whose entries are $g_j(\mathbf{X})$, $j = 1, \ldots, k$ is not singular.

B1 The random variable $(\mathbf{X}, Y)$ verifies
$$v^{lb}(\mathbf{x}) = \mathbf{E}^{lb}\left[\left(\frac{Y - m(\mathbf{X})}{Y}\right)^2 \middle| \mathbf{X} = \mathbf{x}\right] \tag{15}$$
is uniformly integrable with respect to F^{lb}.

Therefore, under condition A1 Proposition 2.1 also holds in this setup, $\mathbf{g}$, $\mathbf{G}$, and $\mathbf{B}$ being defined in the same way as a consequence of the length–bias in the response. Moreover, the definitions we have given for I^{lb}, I^{lb}_n, I^{lb}_n, $R^{lb^1}_n$ and R^{lb}_n are also appropriate in this context substituting univariate variables x, z and x_i by their multivariate counterparts $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{x}_i$, where for vectors $\mathbf{x}$ and $\mathbf{z}$ with components $(x_1, \ldots, x_d)$ and $(z_1, \ldots, z_d)$ respectively, $\mathbf{x} \leq \mathbf{z}$ means that $\mathbf{x} \in (-\infty, z_1] \times \cdots \times (-\infty, z_d]$.

Propositions 2.3 and 2.2 are also true in this context because the covariate being multidimensional carries no change in argumentation taking into account the assumptions A1 and B1. However,in the case of the weak convergence for R^{lb}_n things are more complicated than in the univariate setup because of tightness.

Proposition 3.1 *If assumption B1 is fulfilled:*

$$R^{lb}_n(\mathbf{x}) \to R^{lb}_\infty(\mathbf{x})$$

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

$$\mathbf{Cov}\left[R^{lb}_\infty(\mathbf{x}), R^{lb}_\infty(\mathbf{x}')\right] = \int_\infty^{\mathbf{x}\wedge\mathbf{x}'} v^{lb}(\mathbf{z})\, dF^{lb}(\mathbf{z}).$$

The space $D[\mathbf{R}]^d$ is the natural extension to this setup of $D[-\infty, \infty]$, check Domínguez and Lobato (2003) and the references therein for details. The previous result enables us to provide asymptotics for the statistics we will use to perform goodness of fit because $R^1_n(\mathbf{x})$ admits the same strong uniform representation we give in Proposition 2.5 for the bivariate case. As a consequence we obtain the same kind of asymptotic for this process.

Theorem 3.1 *Under the assumptions A1 and B1, if $g_1, \ldots, g_k$ are uniformly bounded functions in $\mathbf{R}^d$, then:*

$$R^{lb^1}_n(\mathbf{x}) \to R^{lb^1}_\infty(\mathbf{x})$$

in distribution in the space $D[\mathbf{R}]^d$, where $R^{lb^1}_\infty(\mathbf{x})$ is a gaussian process *with null expectation and whose covariance function is given by*

$$\begin{aligned} K^{lb}(\mathbf{x},\mathbf{x}') &= \mathbf{Cov}\left[R_\infty^{lb^1}(\mathbf{x}), R_\infty^{lb^1}(\mathbf{x}')\right] \\ &= \int_\infty^{\mathbf{x}\wedge\mathbf{x}'} v^{lb}(\mathbf{z})\, dF^{lb}(\mathbf{z}) \\ &+ \int_\infty^{\mathbf{x}} v^{lb}(\mathbf{z})\, \mathbf{G}(\mathbf{x}')^T\, \mathbf{L}^{-1}\mathbf{g}(\mathbf{z})\, dF^{lb}(\mathbf{z}) \\ &+ \int_\infty^{\mathbf{x}'} v^{lb}(\mathbf{z})\, \mathbf{G}(\mathbf{x})^T\, \mathbf{L}^{-1}\mathbf{g}(\mathbf{z})\, dF^{lb}(\mathbf{z}) \\ &+ \mathbf{G}(\mathbf{x}')^T\mathbf{L}^{-1}\boldsymbol{\Sigma}^{lb}\mathbf{L}^{-1}\mathbf{G}(\mathbf{x}), \end{aligned}$$

where

$$\boldsymbol{\Sigma}^{lb} = \mathbf{E}^w\left[v^{lb}(\mathbf{X})\mathbf{g}(\mathbf{X})\mathbf{g}(\mathbf{X})^T\right].$$

The asymptotic distribution the previous theorem gives us depends on a relatively complex expression. To avoid this drawback, we propose bootstrap scheme

$$\mathbf{x}_i^* = \mathbf{x}_i;\ y_i^* = \mathbf{g}(\mathbf{x}_i^*)^T\hat{\boldsymbol{\beta}}_n + \hat{\epsilon}_i^*; \hat{\epsilon}_i^* = \hat{\epsilon}_i\,\gamma_i; \tag{16}$$

we use in the bivariate case. Hence, both $R_n^{lb^1*}$ and R_n^{lb*} are defined in a similar fashion.

As a consequence of $\mathbf{x}_i^* = \mathbf{x}_i$, we obtain the strong uniform consistency of the bootstrap estimator for $\hat{\boldsymbol{\beta}}_n^*$. The strong uniform approximation as that presented in Proposition 2.7 follows by means of the same argumentation given in the proof of this result. Therefore, we can prove that the bootstrap process $R_n^{lb^1*}(\mathbf{x})$ converges weakly to the same limit the process the process $R_n^{lb^1}(\mathbf{x})$ has, which shows the consistency of the bootstrap scheme used.

Theorem 3.2 *Under the assumptions made in Theorem 3.1:*

$$R_n^{lb^1*}(\mathbf{x}) \to R_\infty^{lb^1}(\mathbf{x})$$

in distribution in the space $D[\mathbf{R}]^d$, where $R_\infty^{lb^1}(\mathbf{x})$ is a gaussian process with null expectation and whose covariance function is given by $K^{lb}(\mathbf{x},\mathbf{x}')$ as in Theorem 3.1

3.2 Selection bias

As we have explained in Section 1, there are a number of situations from which length–biased data (see Section 2) is only a particular example where the observation of the random phenomena of interests is not directly available. Using the previous discussion and knowledge we have gained with the explanations given in Section 2, we will consider in this section a more general situation that covers a number of selection bias situations present in applications where the observed distribution is given by:

$$dF^w(\mathbf{x}, y) = \frac{w(\mathbf{x}, y)}{\mu_w} dF(\mathbf{x}, y),$$

hence when $w(\mathbf{x}, y) = y$ we have that our data is response length–biased. While we will mainly focus on the theoretical properties needed to implement the goodness of fit test we have been dealing with, the first part of this section will be concerned also with some issues related to the way selection bias affects the estimation.

In the particular case of the estimation of the regression function, equation (2) becomes

$$\mathbf{E}^w \left[Y | \mathbf{X} = x\right] = m(\mathbf{x}) \left(1 + \frac{\mathbf{Cov}\left[Y, w(\mathbf{X}, Y) | \mathbf{X} = \mathbf{x}\right]}{m(\mathbf{x})\mathbf{E}\left[w(\mathbf{X}, Y) | \mathbf{X} = \mathbf{x}\right]} \right), \tag{17}$$

from which we can see how the conditional covariance between the response variable and $w(\mathbf{X}, Y)$ influences the regression estimation.

Again we would like to point out that when data is biased by selection, $\mathbf{X}$ marginal are affected in the following way

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

where $\mu_w = \mathbf{E}\left[w(\mathbf{X}, Y)\right]$, and the Y^w marginal density becomes

$$f_Y^w(y) = \frac{\mathbf{E}\left[w(\mathbf{X}, Y) | Y = y\right]}{\mu_w} f_Y(y).$$

In order to deal with the following examples we will require assumptions A1 and B1, that in this case becomes

$$v^w(\mathbf{x}) = \mathbf{E}^w \left[\left(\frac{Y - m(\mathbf{X})}{w(\mathbf{X}, Y)} \right)^2 \middle| \mathbf{X} = \mathbf{x} \right] \tag{18}$$

is uniformly integrable with respect to F^w. However, besides these two conditions, we will also require that the function w is positive and its support contains the support of the distribution F, that is to say:

A2 The function w verifies

$$w(\mathbf{X}, Y) > 0 \;\; a.s. \;\text{ and }\; \mathrm{supp}(F) \subset \mathrm{supp}(w)$$

where supp(g) is the support of g. Otherwise, it is impossible to recover the whole distribution only using a sample from F^w. But in any case, estimation and inference can be addressed in a common framework, and in this sense, in a unified way.

With respect to the previous setup, the only change we should introduce is that $\mathbf{B}$ is a diagonal matrix with entries $1/w_i$ where $w_i = w(\mathbf{x}_i, y_i)$, but again both $\mathbf{g}$ and $\mathbf{G}$ are defined in the same way, and hence the Proposition 2.1 also holds in this setup as assumption A1 is fulfilled.

As is obvious, the integrated regression function definition does not change at all, but of course the way it is computed using observations from the distribution F^w should be reconsidered. Following the ideas in Section 2 jointly with the previous discussion about the way marginals are modified in this setup, the integrated regression function for F computed with respect to F^w is defined as

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

Now, as in the length–biased case, we obtain

Proposition 3.2 *The function*

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

for $\mathbf{x} \in \mathbf{R}^d$ *is the unique measurable function* F^w*–a.e. such that*

$$I(\mathbf{x}) = \int_{\infty}^{\mathbf{x}} h(\mathbf{z}) \, dF^w(\mathbf{z}).$$

Again following the argumentation given in Section 2, the estimation of I in this context should be carried out using

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

Moreover, it can be proved that:

Proposition 3.3

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

uniformly and almost surely

Naturally, we have that both $R_n^{w^1}$ and R_n^w are defined as

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

that, because of assumptions A1 and B1, have the same kind of asymptotic behavior that we found in the multivariate response length–biased case.

Theorem 3.3 *Under the assumptions A1 and B1, if $g_1, \ldots, g_k$ are uniformly bounded functions in $\mathbf{R}^d$, then:*

$$\begin{array}{rcl} R_n^w(\mathbf{x}) & \rightarrow & R_\infty^w(\mathbf{x}), \\ R_n^{w^1}(\mathbf{x}) & \rightarrow & R_\infty^{w^1}(\mathbf{x}) \end{array}$$

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

$$\mathbf{Cov}\left[R_\infty^w(\mathbf{x}), R_\infty^w(\mathbf{x}')\right] = \int_\infty^{\mathbf{x}\wedge\mathbf{x}'} v^w(\mathbf{z})\, dF^w(\mathbf{z}).$$

The process $R_\infty^{w^1}(\mathbf{x})$ is also a gaussian process with null expectation and whose covariance function is:

$$\begin{aligned} K^w(\mathbf{x}, \mathbf{x}') &= \mathbf{Cov}\left[R_\infty^{w^1}(\mathbf{x}), R_\infty^{w^1}(\mathbf{x}')\right] \\ &= \int_\infty^{\mathbf{x}\wedge\mathbf{x}'} v^w(\mathbf{z})\, dF^w(\mathbf{z}) \\ &+ \int_\infty^{\mathbf{x}} v^w(\mathbf{z})\, \mathbf{G}(\mathbf{x}')^T \mathbf{L}^{-1}\mathbf{g}(\mathbf{z})\, dF^w(\mathbf{z}) \\ &+ \int_\infty^{\mathbf{x}'} v^w(\mathbf{z})\, \mathbf{G}(\mathbf{x})^T \mathbf{L}^{-1}\mathbf{g}(\mathbf{z})\, dF^w(\mathbf{z}) \\ &+ \mathbf{G}(\mathbf{x}')^T\mathbf{L}^{-1}\mathbf{\Sigma}^w\mathbf{L}^{-1}\mathbf{G}(\mathbf{x}), \end{aligned}$$

where

$$\mathbf{\Sigma}^w = \mathbf{E}^w\left[v^w(\mathbf{X})\mathbf{g}(\mathbf{X})\mathbf{g}(\mathbf{X})^T\right].$$

As we can see, these asymptotics are rather complicated, therefore the same motivation that led us to consider the use of bootstrap in the length–biased case persists. In this case, the bootstrap scheme we propose is the same as that given in the length–biased case but using the reciprocal of the w_i

$$\mathbf{x}_i^* = \mathbf{x}_i;\; y_i^* = \mathbf{g}(\mathbf{x}_i^*)^T\hat{\boldsymbol{\beta}}_n + \hat{\epsilon}_i^*; \hat{\epsilon}_i^* = \hat{\epsilon}_i\,\gamma_i;$$

and both $R_n^{w^1*}$ and R_n^{w*} are defined in a similar way:

$$R_n^{w^1*}(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{w_i}\Big(y_i^* - \mathbf{g}(\mathbf{x}_i^*)^T\hat{\boldsymbol{\beta}}_n^*\Big)\mathbf{1}_{\{\mathbf{x}_i^*\leq\mathbf{x}\}},$$

while the asymptotic behavior in this case takes into account the limit process $R_\infty^{w^1}$.

Theorem 3.4 *Under the assumptions made in Theorem 3.3:*

$$R_n^{w^1*}(\mathbf{x}) \to R_\infty^{w^*}(\mathbf{x})$$

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

Therefore the bootstrap scheme we have introduced is also useful to calibrate rejection region boundaries.

3.3 Some Examples

Let us consider now several different examples of selection bias from a more detailed perspective. We will focus on the general effect selection bias produces on the observed distributional behaviour taking into account both marginals, joint distributions and the regression function.

Example 3.1 ***Length–biased data***

As we have seen, in this case, $w(\mathbf{x}, y) = y$ and, the regression function of the observed data is given by

$$\mathbf{E}^w[Y|\mathbf{X} = \mathbf{x}] = m(\mathbf{x})\left(1 + \frac{\sigma^2(\mathbf{x})}{m(\mathbf{x})^2}\right).$$

Hence, the usual regression estimation can not be used in this case without a correction or a *compensation* that avoids the bias present in the data. But also notice that in this case, $f_{\mathbf{X}}^w(\mathbf{x}) = m(\mathbf{x})\mu_X^{-1} f_{\mathbf{X}}(x)$, so the marginals of the covariates are also affected.

It is of interest to mention, when we are dealing with length–biased data in one of the covariate components of the vector $\mathbf{X} = (X_1 \ldots, X_j)$, that is to say when $w(\mathbf{x}, y) = x_j$, we have that

$$\mathbf{Cov}[Y, w(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}] = \mathbf{Cov}[Y, X_j|\mathbf{X} = \mathbf{x}] = 0$$

hence, the regression function of the observed data agrees with $m(\mathbf{x})$ as in this case

$$\mathbf{E}^w[Y|\mathbf{X} = \mathbf{x}] = m(\mathbf{x})\left(1 + \frac{0}{x_j\ m(\mathbf{x})}\right)$$

but, $f_X^w(\mathbf{x}) = \mu_{X_j}^{-1} x_j f_X(\mathbf{x})$, where $\mu_{X_j} = \mathbf{E}^w[X_j]$. Intuitively, this has to do with the fact that conditioning on covariates, and hence the regression function, does not depend on covariates marginals. Notice that this means that as a consequence of the length–bias in the covariates, marginals may change but the regression function remains the same. This also happens when the selection bias depends only on covariates, that is to say, when $w(\mathbf{x}, x) = h(\mathbf{x})$ for a given function h. As a consequence, in those cases where the length–bias depends on the covariates, it is not necessary to correct or to compensate the bias present in the data.

Let us now focus on other kinds of selection bias related to lifetime analysis, see for exmple Ansell and Phillips (1994). In some practical situations where the analysis of the duration of events is involved (mainly in reliability or epidemiological contexts) there sometimes exists the need to estimate the running time of certain events that are somehow related. For example, the lifetime of a system made with different components (subsystems) depends on the running time of its components. In fact, these subsystems keep on running according to a certain policy that defines the running time of the system depending on the lifetime of its different parts. Now, depending on the way these times are registered, you may not have a sample from the real durations of the events you are interested in, but a sample from a random variable that is somehow related to it. For the shake of ease in exposition, let us assume first that these systems have only two components with random lifetimes X and Y respectively, that are possibly related and that, for whatever reason, we are interested in the relationship between the random lifetime of the subsystems X and Y.

The usual way to estimate characteristics of the lifetime of any of the components of the system is to perform some kind of experiment in which n of these systems are set to work together, later measuring the running time $Z = w(X, Y)$ of each of them, jointly with the durations of the components X and Y. In this way, what we obtain is an i.i.d. sample from the random population (X, Y). Nevertheless, there exists another different way in which we can obtain information from (X, Y). We can also think about taking a sample at a given point in time while the systems are running; in this case those systems with larger lifetimes will be more likely to be in the sample. Therefore, in the latter case, while the lifetimes of these two subsystems are distributed as $dF(x, y)$, the observed lifetimes of the two components (X^w, Y^w) are distributed according to

$$dF(x, y) = \frac{w(x, y)}{\mu_w} dF(x, y)$$

with $w(x, y)$ depending on the policy the system uses to manage their components. We will now consider different policies.

Example 3.2 ***Replacement of subsystem policy***

In this case, one of the subsystems start to run when the one that is running stops. Hence, the system lifetime amounts to the sum of the lifetimes of the two components as they run one after the other and $Z = X + Y$. As a consequence

$$\mathbf{Cov}\left[Y, X+Y|X=x\right] = \mathbf{Var}\left[Y|X=x\right],$$

and the estimation of the regression function with respect to the observed distribution is then

$$\mathbf{E}^w\left[Y|X=x\right] = m(x)\left(1 + \frac{\mathbf{Var}\left[Y|X=x\right]}{m(x)(m(x)+x)}\right). \tag{19}$$

Recall that $w(X,Y) = X + Y > 0$ a.s., which is accomplished when both lifetimes are positive with probability one. The marginal densities are also modified, being proportional to $y + \mathbf{E}\left[X|Y=y\right]$ in the case of Y^w and to $x + m(x)$ in the case of X^w.

Notice that in the multivariate setup $(\mathbf{X}, Y)$, when $Z = Y + \sum_{j=1}^d X_j$, equation (19) should be replaced by

$$\mathbf{E}^w\left[Y|X=x\right] = m(x)\left(1 + \frac{\mathbf{Var}\left[Y|X=x\right]}{m(x)\Big(m(x)+\sum_{j=1}^d x_j\Big)}\right).$$

Example 3.3 ***Running in parallel policy***

Now subsystems works altogether from the beginning, and the system keeps on running till the last component fails to run. Hence, the system lifetime depends on the lifetime of the longer lasting component, and its lifetime is given by $Z = \max(X,Y)$. As a consequence, if we observe the lifetime phenomena at a definite point of time, we make observations whose probability of being registered in the sample is proportional to the lifetime of the longer lasting component.

Notice first that

$$\mu_{\max}(x) = \mathbf{E}\left[\max(X,Y)|X=x\right] = xF_{Y|X=x}(x) + \overline{G^1_{Y|X=x}}(x).$$

where following and generalizing some expressions in Patil and Taillie (1989), $\overline{G^k_W}(x)$ stands for

$$\overline{G^k_W}(x) = \int_x^\infty w^k f_W(w)\, dw,$$

being f_W the density of a random variable W, and

$$G_W^k(x) = \int_0^x w^k f_W(w)\, dw,$$

These functions are a kind of accumulation of the kth–order moment of W from the value of x or up to the value of x respectively. Hence,

$$\overline{G_W^k}(x) = \mathbf{E}\left[W^k\right] - G_W^k(x),$$

that in the particular case of $W = Y|X = x$ means

$$\overline{G_{Y|X=x}^k}(x) = \mathbf{E}\left[Y^k|X = x\right] - G_{Y|X=x}^k(x),$$

moreover $G_W^0(x) = F_W(x)$.

Now with this notation, and as far as

$$\begin{aligned}&\mathbf{Cov}\left[Y, \max(X,Y)|X = x\right] = \\ &\qquad x\mathbf{Cov}\left[Y, \mathbf{1}_{\{X>Y\}}|X = x\right] + \mathbf{Cov}\left[Y, Y\mathbf{1}_{\{X\leq Y\}}|X = x\right]\end{aligned}$$

the regression with respect to the observed distribution is then

$$\mathbf{E}^w\left[Y|X = x\right] = m(x)\left(1 + \frac{xG_{Y|X=x}^1(x) + \overline{G_{Y|X=x}^2}(x) - m(x)\ \mu_{\max}(x)}{m(x)\ \mu_{\max}(x)}\right),$$

which shows us the complexity of the regression estimation bias term, telling us about how difficult its correction can be.

Recall that in this case the marginal densities can be heavily modified. In the case of Y^w the marginal density is now proportional to

$$yF_{X|Y=y}(y) + \overline{G_{X|Y=y}^1}(y),$$

while the marginal density for X^w is proportional to $\mu_{\max}(x)$. As we can see, in both cases, the relationship with the marginal of the original random variables can be rather complicated.

Example 3.4 ***Running in series policy***

Again the system lifetime depends on the lifetime of components running at the same time. But in this case, as each of the subsystem requires the work of the previous subsystem done before it starts to work, the whole system lifetime last till the first component fails to run, that is to say $Z = \min(X, Y)$. So if we observe the lifetime phenomena at a definite point in time, the chance of the observation being in the sample is now proportional to the lifetime of the first component that fails.

In the case of the regression function, and using previous notation we have that if

$$\mu_{\min}(x) = \mathbf{E}\left[\min(X,Y)|X=x\right] = x\overline{F_{Y|X=x}}(x) + G^1_{Y|X=x}(x),$$

the regression with respect to the observed distribution is then

$$\mathbf{E}^w\left[Y|X=x\right] = m(x)\left(1+\frac{x\overline{G^1_{Y|X=x}}(x) + G^2_{Y|X=x}(x) - m(x)\ \mu_{\min}(x)}{m(x)\ \mu_{\min}(x)}\right)$$

as a consequence of being

$$\begin{aligned}&\mathbf{Cov}\left[Y,\min(X,Y)|X=x\right] = \\ &\qquad x\mathbf{Cov}\left[Y,\mathbf{1}_{\{X\leq Y\}}|X=x\right] + \mathbf{Cov}\left[Y,Y\mathbf{1}_{\{X>Y\}}|X=x\right],\end{aligned}$$

showing us again the complexity of the correction of the regression estimation bias in this particular case.

The marginal densities are now proportional to

$$y\overline{F_{X|Y=y}}(y) + G^1_{X|Y=y}(y),$$

in the case of Y^w and to $\mu_{\min}(x)$ in the case of X^w.

It should also be mentioned that if instead of a bivariate random variable (X,Y) the phenomena we are interested in is distributed according to a multivariate distribution $(\mathbf{X},Y)$, then we need to modify $\mu_{\max}(x)$ and $\mu_{\min}(x)$ in the following way:

$$\begin{aligned}\mu_{\max}(\mathbf{x}) &= \mathbf{E}\left[\max(\mathbf{X},Y)|\mathbf{X}=\mathbf{x}\right] \\ &= \max(\mathbf{x})F_{Y|\mathbf{X}=\mathbf{x}}(\max(\mathbf{x})) + \overline{G^1_{Y|\mathbf{X}=\mathbf{x}}}(\max(\mathbf{x})), \\ \mu_{\min}(\mathbf{x}) &= \mathbf{E}\left[\min(\mathbf{X},Y)|\mathbf{X}=\mathbf{x}\right] \\ &= \min(\mathbf{x})\overline{F_{Y|\mathbf{X}=\mathbf{x}}}(\min(\mathbf{x})) + G^1_{Y|\mathbf{X}=\mathbf{x}}(\min(\mathbf{x})).\end{aligned}$$

It is worth pointing out that these kinds of selection bias can also appear in some epidemiological or disease studies where the duration of the evolution of illness is measured by means of encountering the patients and/or using screening methods. We now turn to a different framework that appears frequently in a number of different research areas.

Example 3.5 ***Stratified sampling***

Stratified sampling is characterized by the fact that some subsets of the population are sampled with a given artificial probability. Assuming that our population is a bivariate random variable $(\mathbf{X}, Y)$ in $\mathbf{R}^{d+1}$ and that the measurable set $A \subset \mathbf{R}^{d+1}$ is sampled with probability $p_A^w > 0$, the observed bivariate population $(\mathbf{X}, Y)$ is distributed according to

$$dF^w(\mathbf{x}, y) = w_A(\mathbf{x}, y) dF(\mathbf{x}, y)$$

where

$$w_A(\mathbf{x}, y) = \frac{p_A^w}{p_A} \mathbf{1}_{\{(\mathbf{x},y) \in A\}} + \frac{1 - p_A^w}{1 - p_A} \mathbf{1}_{\{(\mathbf{x},y) \in A^c\}}$$

A^c being the $\mathbf{R}^{d+1}$ complement of the set A and p_A the probability of having an observation of the population $(\mathbf{X}, Y)$ in A, that is to say $p_A = \mathbf{P}((\mathbf{X}, Y) \in A)$ that should be positive.

Moreover, because of begin dealing with regression function, let us introduce $p_B(\mathbf{x})$ and $m_B(\mathbf{x})$ defined for a measurable set $B \in \mathbf{R}^{d+1}$ as $\mathbf{P}((\mathbf{X}, Y) \in B | \mathbf{X} = \mathbf{x})$ and $\mathbf{E}\left[Y \mathbf{1}_{\{(\mathbf{X},Y) \in B\}} | \mathbf{X} = \mathbf{x}\right]$ respectively.

Notice also that

$$w_A(\mathbf{x}, y)^{-1} = \frac{p_A}{p_A^w} \mathbf{1}_{\{(\mathbf{x},y) \in A\}} + \frac{1 - p_A}{1 - p_A^w} \mathbf{1}_{\{(\mathbf{x},y) \in A^c\}},$$

and the fact that, while $\mathbf{E}\left[w_A(\mathbf{X}, Y)\right] = 1$,

$$\mathbf{E}\left[w_A(\mathbf{X}, Y) | \mathbf{X} = \mathbf{x}\right] = \frac{p_A^w}{p_A} p_A(\mathbf{x}) + \frac{1 - p_A^w}{1 - p_A} p_{A^c}(\mathbf{x}).$$

This leads to the following relationship

$$\mathbf{E}^w\left[Y | \mathbf{X} = \mathbf{x}\right] = m(\mathbf{x}) \left(1 + \frac{m_A(\mathbf{x})}{m(\mathbf{x}) p_A(\mathbf{x})} + \frac{m_{A^c}(\mathbf{x})}{m(\mathbf{x}) p_{A^c}(\mathbf{x})} - 1 \right)$$

where we assumed that $m_B(\mathbf{x}) p_B(\mathbf{x})^{-1}$ is null whenever $p_B(\mathbf{x}) = 0$.

The marginals are proportional to

$$\frac{p_A}{p_A^w} \mathbf{P}\{(\mathbf{X}, Y) \in A | X_j = x\} + \frac{1 - p_A}{1 - p_A^w} \mathbf{P}\{(\mathbf{X}, Y) \in A | X_j = x\}$$

in the case of X_j^w, and to

$$\frac{p_A}{p_A^w} \mathbf{P}\{(\mathbf{X}, Y) \in A | Y = y\} + \frac{1 - p_A}{1 - p_A^w} \mathbf{P}\{(\mathbf{X}, Y) \in A^c | Y = y\}$$

in the case of Y^w. All these expressions make clear that the occurrence of the $(\mathbf{X}, Y) \in A$ given that $X_j = x$ and/or $Y = y$ plays a major role on the effect stratification has on the observed sample.

It is worth mentioning, that A can be any interval or union of intervals in $\mathbf{R}^{d+1}$, in particular those intervals of the form $\mathbf{R}^d \times [a, b]$. Regarding this issue, notice that when stratum selection is based only on covariates, that is to say $A = B \times \mathbf{R}$ with $B \in \mathbf{R}^d$, it does not affect the regression estimation. It is also worth noticing the importance in these expressions of the relative quotient between p_A and p_A^w, and their complementary counterparts. It particular, notice these values can change dramatically both the true regression function.

4 Some Simulations

The simulations we will carry out in this section will be devoted to the analysis of the performance of the test introduced in Section 2 in the length–bias framework. The analysis will be mainly focused on the acceptance/rejection performance. The examples we will show to perform the simulation are based on models from the simulations in Stute et al. (1998), with suitable modifications to obtain a positive response in all the examples.

Basically, if the population we are interested in is distributed according to $(\mathbf{X}, Y)$ the modifications we introduce consist in the use of an Exponential $\mathcal{E}(\sigma)$ random variable as the additive random error, or in the use of suitable multiplicative error. In the first case, we have $Y = m(\mathbf{X}) + \mathcal{E}(\sigma)$, hence $Y = m(\mathbf{X}) + \sigma + \epsilon$, where ϵ has null expectation and variance equal to σ^2. Multiplicative error models are defined as $Y = m(\mathbf{X})\big(1 + \sigma\,\mathcal{U}\big(-\sqrt{3}, \sqrt{3}\big)\big)$, where $\mathcal{U}(a, b)$ is a uniform random variate in $[a, b]$, therefore these models lead to a positive response for $\sigma < 1/\sqrt{3}$. With the aim of better appreciating the behavior of these tests under different conditions an additive perturbation term of the regression function under the null hypotheses has been introduced. The perturbation term is proportional to A.

For each of the first three examples considered, we have added figures that show the plot of the power function for both statistics (KnpVal for $K_n^\infty$and WnpVal for W_n^2) when $\sigma = 0.1, 0.5$ and the sample sizes n are $50, 100, 200$, the bootstrap resampling size B is 400, and A ranges from -2.0 to 2.0. Besides the plot, for each of the examples we have added tables which present the rejection rates under different conditions: $\sigma = 0.1, 0.5$, $n = 50, 100, 200$ $\alpha = 0.05, 0.01$, being $B = 400$ again.

Example 4.1 ***Model I in Stute et al. (1998) with $\mathcal{E}(\sigma)$ errors*** In this case we have considered (X, Y) according to

$$Y = 5X + AX^2 + \mathcal{E}(\sigma), \quad X \sim \mathcal{U}(0, 1),$$

so for $A = 0$ the model is linear in the covariate, while in the case of $A \neq 0$ we have that the null hypothesis is no longer true for different degrees of separation from the linear dependence on covariate.

Example 4.2 ***Model I in Stute et al. (1998) with multiplicative errors*** While the regression function is the same as in the previous example, in this particular case, (X, Y) are related according to

$$Y = \left(5X + AX^2\right)\Big(1 + \sigma\,\mathcal{U}\Big(-\sqrt{3}, \sqrt{3}\Big)\Big), \quad X \sim \mathcal{U}(0,1),$$

which changes in a noticeable manner the distributional behavior. Again different values of A control the degree of separation of the model from the linearity in the covariate. However, in this case it is interesting to point out that the error variance also includes the regression function. This decreases the power function in a noticeable manner for some values of A, especially in the case of $\sigma = 0.5$.

Example 4.3 ***Model III in Stute et al. (1998) with multiplicative errors*** This is a multivariate regression problem in which the population $(\mathbf{X}, Y)$ are related by means of

$$\begin{aligned} Y &= (2 + 5X_1 - 2X_2 + AX_1X_2)\Big(1 + \sigma\,\mathcal{U}\Big(-\sqrt{3}, \sqrt{3}\Big)\Big), \\ X_1, X_2 &\sim \mathcal{U}(0,1) \times \mathcal{U}(0,1), \end{aligned}$$

Notice that, in this particular example, A controls the degree of separation of the model from the linearity but by means of an interrelationship between both covariates. Again in this case it is worth mentioning how multiplicative error affects the performance of the testing power.

As we can see, most of the power function plots exhibit the expected behavior: the percentage of rejections increases as the absolute value of A increases. It is worth mentioning that, as was pointed out in Stute (1997) for the usual sampling framework, the lack of power is especially noticeable when σ is large. It is also interesting to point out that in the case of example 4.2, the power function plot exhibits some kind of asymmetry. This lack of symmetry may be related to the lack of orthogonality between the regression function and the perturbation, and it may be empowered by the multiplicative errors, as in this particular case, variance is proportional to the regression function, which is also affected by the perturbation term.

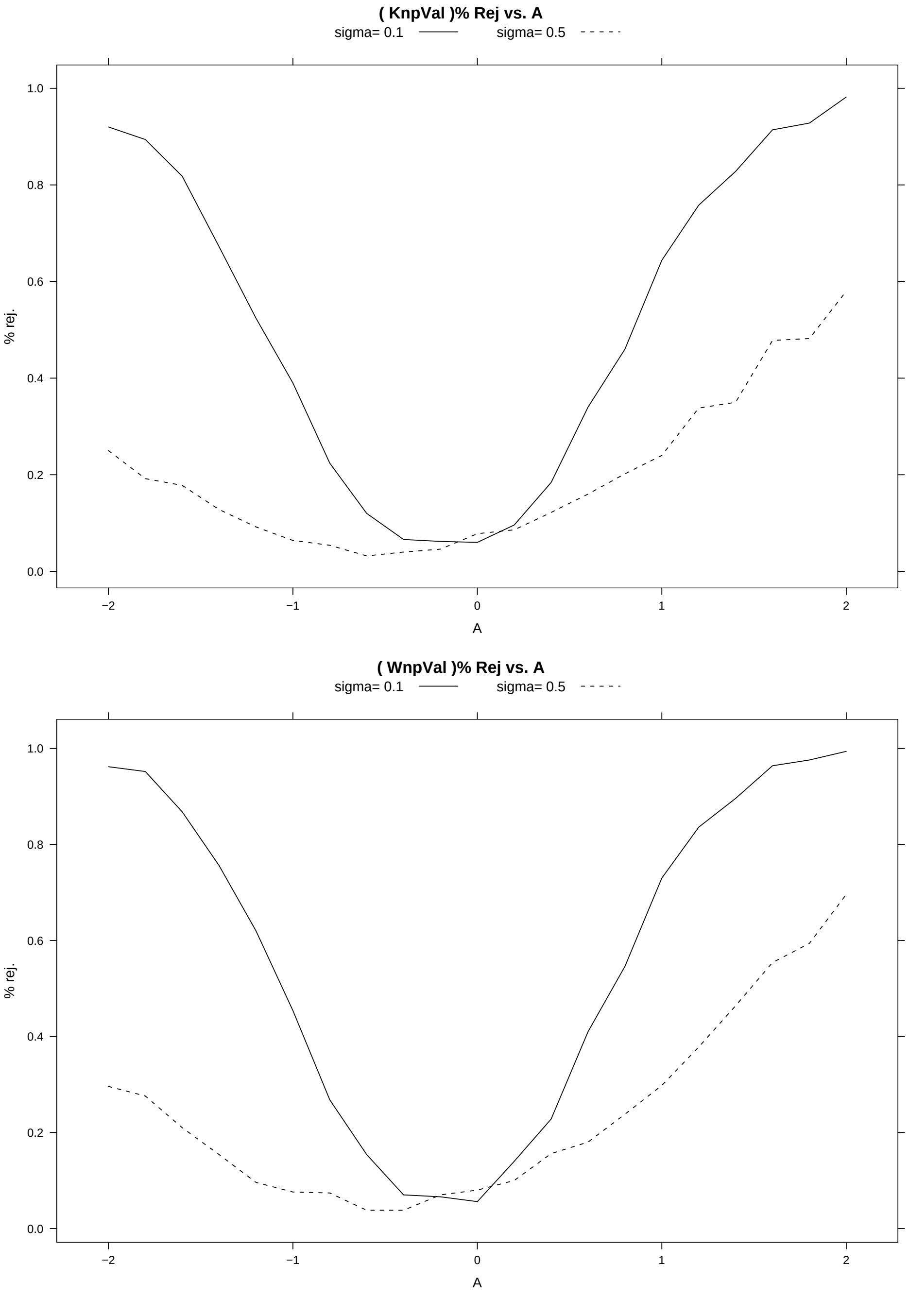

Figure 1: Power function for example 4.1($\alpha = 0.05$).

Table 1: Rejection percentage of H_0 for K_n^∞ depending on A for example 4.1.

α	n	A	σ	K_n^∞	α	n	A	σ	K_n^∞
0.01	50	-1	0.1	0.024	0.05	50	-1	0.1	0.114
		0		0.004			0		0.036
		1		0.050			1		0.190
		-1	0.5	0.008			-1	0.5	0.030
		0		0.002			0		0.022
		1		0.022			1		0.088
	100	-1	0.1	0.048		100	-1	0.1	0.188
		0		0.004			0		0.060
		1		0.124			1		0.350
		-1	0.5	0.004			-1	0.5	0.046
		0		0.008			0		0.064
		1		0.042			1		0.116
	200	-1	0.1	0.186		200	-1	0.1	0.404
		0		0.008			0		0.062
		1		0.318			1		0.616
		-1	0.5	0.010			-1	0.5	0.054
		0		0.008			0		0.072
		1		0.090			1		0.256

Table 2: Rejection percentage of H_0 for W_n^2 depending on A for example 4.1.

α	n	A	σ	W_n^2	α	n	A	σ	W_n^2
0.01	50	-1	0.1	0.034	0.05	50	-1	0.1	0.142
		0		0.002			0		0.054
		1		0.066			1		0.240
		-1	0.5	0.010			-1	0.5	0.042
		0		0.004			0		0.036
		1		0.034			1		0.106
	100	-1	0.1	0.074		100	-1	0.1	0.258
		0		0.008			0		0.056
		1		0.176			1		0.428
		-1	0.5	0.012			-1	0.5	0.052
		0		0.006			0		0.070
		1		0.044			1		0.168
	200	-1	0.1	0.252		200	-1	0.1	0.484
		0		0.012			0		0.056
		1		0.464			1		0.738
		-1	0.5	0.020			-1	0.5	0.074
		0		0.026			0		0.086
		1		0.140			1		0.284

Table 3: Rejection percentage of H_0 for K_n^∞ depending on A for example 4.2.

α	n	A	σ	K_n^∞	α	n	A	σ	K_n^∞
0.01	50	-1	0.1	0.040	0.05	50	-1	0.1	0.168
		0		0.012			0		0.046
		1		0.030			1		0.092
		-1	0.5	0.014			-1	0.5	0.056
		0		0.006			0		0.062
		1		0.020			1		0.076
	100	-1	0.1	0.112		100	-1	0.1	0.304
		0		0.016			0		0.056
		1		0.040			1		0.148
		-1	0.5	0.010			-1	0.5	0.048
		0		0.022			0		0.058
		1		0.006			1		0.060
	200	-1	0.1	0.240		200	-1	0.1	0.512
		0		0.010			0		0.052
		1		0.088			1		0.240
		-1	0.5	0.008			-1	0.5	0.048
		0		0.002			0		0.044
		1		0.018			1		0.068

All the tables are similar regarding the behavior. When $A = 0$, the rejection rate agrees more or less with the nominal level α, and increases when $A \neq 0$. The effect of $\sigma = 0.5$ is clearly noticeable as the rejection rate increase is not as large as in the case of $\sigma = 0.1$.

Acknowledgment

This work was partially supported by DGI Grants MTM2005-01464 and MTM2005-00820 (european FEDER support included), jointly with funds from the Consolidated Group DGA 2005–S11 and the DGA(CONSI+D) and CAI program EUROPA.

Table 4: Rejection percentage of H_0 for W_n^2 depending on A for example 4.2.

α	n	A	σ	W_n^2	α	n	A	σ	W_n^2
0.01	50	-1	0.1	0.054	0.05	50	-1	0.1	0.218
		0		0.010			0		0.048
		1		0.032			1		0.102
		-1	0.5	0.018			-1	0.5	0.064
		0		0.008			0		0.058
		1		0.024			1		0.074
	100	-1	0.1	0.180		100	-1	0.1	0.392
		0		0.008			0		0.042
		1		0.058			1		0.186
		-1	0.5	0.010			-1	0.5	0.066
		0		0.018			0		0.060
		1		0.006			1		0.046
	200	-1	0.1	0.388		200	-1	0.1	0.670
		0		0.020			0		0.044
		1		0.152			1		0.356
		-1	0.5	0.014			-1	0.5	0.042
		0		0.010			0		0.036
		1		0.014			1		0.070

5 Appendix.

5.1 Response Length Biased Data

Proof of Proposition 2.1: Equation (7) can be expressed in matrix terms as

$$\Phi(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{G}\boldsymbol{\beta})^T \mathbf{B} (\mathbf{Y} - \mathbf{G}\boldsymbol{\beta}).$$

The value $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ that maximize this expression is given by

$$0 = -\mathbf{G}^T \mathbf{B} \mathbf{Y} + \mathbf{G}^T \mathbf{B} \mathbf{G} \hat{\boldsymbol{\beta}}_n.$$

Notice that $n^{-1}\mathbf{G}^T\mathbf{B}\mathbf{G}$ is a matrix whose element (j, l) is

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i} g_j(x_i) g_l(x_i),$$

and that

$$\mathbf{E}^{lb} \left[\frac{1}{Y} g_j(X) g_l(X) \right] = \frac{1}{\mu_Y} \mathbf{E} \left[g_j(X) g_l(X) \right].$$

Table 5: Rejection percentage of H_0 for K_n^∞ depending on A for example 4.3.

α	n	A	σ	K_n^∞	α	n	A	σ	K_n^∞
0.01	50	-1	0.1	0.004	0.05	50	-1	0.1	0.054
		0		0.006			0		0.064
		1		0.012			1		
		-1	0.5	0.008			-1	0.5	0.054
		0		0.010			0		0.040
		1		0.018			1		0.084
	100	-1	0.1	0.024		100	-1	0.1	0.082
		0		0.012			0		0.050
		1		0.016			1		0.056
		-1	0.5	0.010			-1	0.5	0.052
		0		0.010			0		0.044
		1		0.002			1		0.034
	200	-1	0.1	0.038		200	-1	0.1	0.154
		0		0.014			0		0.050
		1		0.020			1		0.118
		-1	0.5	0.008			-1	0.5	0.046
		0		0.008			0		0.036
		1		0.012			1		0.052

As all these entries have finite second order moment, the application of the Law of the Iterated Logarithm gives that

$$\frac{1}{n}\mathbf{G}^T\mathbf{B}\mathbf{G} = \frac{1}{\mu_Y}\mathbf{L} + O\left(\sqrt{\frac{\log\log n}{n}}\right) \tag{20}$$

almost surely where $\mathbf{L}$ is given in hypotheses B2. Hence, for a sufficiently large n, we know $\mathbf{G}^T\mathbf{B}\mathbf{G}$ is a non–singular matrix and, as a consequence of writing y_i as $\mathbf{g}(x_i)^T\boldsymbol{\beta}_0+\epsilon_i$, we have that $\mathbf{Y} = \mathbf{G}^T\boldsymbol{\beta}_0+\boldsymbol{\epsilon}$, and therefore:

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

This expression leads to the following almost sure representation of $\hat{\boldsymbol{\beta}}_n$:

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

Now, as $n^{-1}\mathbf{G}^T\mathbf{B}\boldsymbol{\epsilon}$ is a vector with entries given by

$$\frac{1}{n}\sum_{i=1}^{n} g_j(x_i)\frac{\epsilon_i}{y_i},$$

Table 6: Rejection percentage of H_0 for W_n^2 depending on A for example 4.3.

α	n	A	σ	W_n^2	α	n	A	σ	W_n^2
0.01	50	-1	0.1	0.014	0.05	50	-1	0.1	0.088
		0		0.008			0		0.042
		1		0.012			1		0.074
		-1	0.5	0.006			-1	0.5	0.056
		0		0.002			0		0.040
		1		0.012			1		0.066
	100	-1	0.1	0.042		100	-1	0.1	0.162
		0		0.002			0		0.064
		1		0.018			1		0.118
		-1	0.5	0.006			-1	0.5	0.052
		0		0.004			0		0.054
		1		0.008			1		0.052
	200	-1	0.1	0.176		200	-1	0.1	0.430
		0		0.006			0		0.038
		1		0.114			1		0.322
		-1	0.5	0.016			-1	0.5	0.068
		0		0.012			0		0.052
		1		0.012			1		0.068

the application, once more, of the Law of the Iterated Logarithm to each of these entries means that $\mathbf{G}^T\mathbf{B}\boldsymbol{\epsilon}$ is a matrix whose elements are quantities of order $\sqrt{\log\log n/n}$ almost surely and, hence we obtain the almost sure representations given in the proposition. ◇

Proof of Proposition 2.2: This is a consequence of the way we should compute the integrated regression function in this context.

If we assume that there exist a function h such that $I^{lb}(x) = \int_\infty^x h(z)\,dF^{lb}(z)$, then

$$0 = \int_\infty^x (h(z) - \mu_Y)\,dF^{lb}(z)$$

because of the definition for $I^{lb}(x)$, hence $h(z) = \mu_Y$ F^{lb}–a.e. ◇

Proof of Proposition 2.3: $I_n^{lb}(x)$ can be written as an empirical process in terms of the empirical distribution of the observed sample $F_n^{lb}(z)$ in the following way

$$I_n^{lb}(x) = \overline{y}^H \int \mathbf{1}_{\{z\leq x\}}\, dF_n^{lb}(z) = \overline{y}^H \;\; F_n^{lb}\,\mathbf{1}_{\{z\leq x\}}\,,$$

where $F_n^{lb} f$ is used to denote $\int f(z)\, dF^{lb}(z)$. Therefore $I_n^{lb}(x)/\overline{y}^H$ is an empirical process indexed by the following class of functions $\mathcal{C} = \{f_x(z,y) : f_x(z,y) = \mathbf{1}_{\{z \le x\}}\}$, that are F^{lb} measurable VC–classes of functions (they are indicators of semi-axes in $\mathbf{R}$) whose envelope $e_{\mathcal{C}}(z,y)$ is bounded by a constant, therefore $F^{lb}\, e_{\mathcal{C}} < \infty$ and verifying Glivenko–Cantelly property as stated in van der Vaart and Wellner (1996).

The Law of the Iterated Logarithm for the reciprocal of the responses $1/y_i$ proves that $\overline{y}^H - \mu_Y$ is an $O\Big(\sqrt{\log\log n/n}\Big)$ quantity with probability one, and the result follows. $\diamond$

Proof of Proposition 2.4: Recall that the paths of $R_n^{lb}(x)$ have left hand limits and that are continuous on the right hand for every $x \in \mathbf{R}$. We will follow Billingsley (1968) and prove the weak convergence in two steps.The first step is to check that finite–dimensional distribution of $R_n^{lb}(x)$ converges to those of $R_\infty(x)$ and in the second we will deal with tightness.

First, the finite–dimensional distribution convergence. The expectation of $R_n^{lb}(x)$ is null for every x because

$$\mathbf{E}^{lb}\left[\left(\frac{y_i - m(x_i)}{y_i}\right)\Big| x_i\right] = \mathbf{E}\left[y_i - m(x_i) | x_i\right] = 0.$$

Now for the covariance function notice:

$$\begin{aligned} R_n^{lb}(x) R_n^{lb}(x') &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\epsilon_i}{y_i}\right)^2 \mathbf{1}_{\{x_i \le x\}}\mathbf{1}_{\{x_i \le x'\}} \\ &+ \frac{1}{n}\sum_{j \ne i}^{n}\left(\frac{\epsilon_i}{y_i}\right)\left(\frac{\epsilon_j}{y_j}\right)\mathbf{1}_{\{x_i \le x\}}\mathbf{1}_{\{x_j \le x\}} \\ &= A + B. \end{aligned}$$

While from the previous reasoning it is clear that the expectation for B is null, in the case of A we have:

$$\begin{aligned} \mathbf{E}^{lb}\left[R_n^{lb}(x) R_n^{lb}(x')\right] &= \mathbf{E}^{lb}\left[\left(\frac{y_i - m(x_i)}{y_i}\right)^2 \mathbf{1}_{\{x_i \le x \wedge x'\}}\right] \\ &= \mathbf{E}^{lb}\left[v^{lb}(x_i)\mathbf{1}_{\{x_i \le x \wedge x'\}}\right]. \end{aligned}$$

And the finite dimensional distribution convergence follows from the application of the Cramer–World Device or using the multivariate CLT.

For the tightness let us consider the quantile transformed process $Q_n^{lb}(u)$

$$R_n^{lb}(x) = Q_n^{lb}\Big(F^{lb}(x)\Big)$$

where

$$Q_n^{lb}(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{y_i} \Big(y_i - m\Big(F^{lb^{-1}}(u_i) \Big) \Big) \mathbf{1}_{\{u_i \le u\}},$$

and $u_i = F^{lb}(x_i)$, where in this case, we use $F^{lb}(x)$ to denote X^{lb} marginal distribution. Because of this transformation, let us use $\mathbf{E}^q [\cdot]$ to denote the expectation with respect to the distribution of $\big(U, Y^{lb}\big)$ for $U = F^{lb}(X^{lb})$. Then, according to Theorem 15.6 in Billingsley (1968) we have to check that for $0 \le u_1 < u < u_2 \le 1$,

$$\mathbf{E}^q \left[\Big| Q_n^{lb}(u_2) - Q_n^{lb}(u) \Big|^2 \Big| Q_n^{lb}(u) - Q_n^{lb}(u_1) \Big|^2 \right] \le (H(u_2) - H(u_1))^{2\alpha} \tag{21}$$

for a non decreasing continuous function H, whenever $\alpha > 1/2$. Hence if we take

$$\alpha_i\big(u'', u'\big) = \frac{\epsilon_i}{y_i} \mathbf{1}_{\{u' \le u_i \le u''\}},$$

we have that for $0 \le u' < u \le 1$:

$$Q_n^{lb}\big(u''\big) - Q_n^{lb}\big(u'\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\epsilon_i}{y_i} \mathbf{1}_{\{u' \le u_i \le u''\}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha_i\big(u'', u'\big),$$

and the left term in the inequality given in equation (21) is then

$$C = \frac{1}{n^2} \, \mathbf{E}^q \left[\left(\sum_{i=1}^{n} \alpha_i(u_2, u) \right)^2 \left(\sum_{i=1}^{n} \alpha_i(u, u_1) \right)^2 \right]$$

which, as a consequence of each of the terms in both sums having null expectation, can be bounded using Lemma 5.1 in Stute (1997) with $\alpha_i = \alpha_i(u_2, u)$ and $\beta_i = \alpha_i(u, u_1)$. Therefore

$$C \le \frac{1}{n^2} 3n(n-1) \mathbf{E}^q \left[\alpha_1(u_2, u)^2 \right] \mathbf{E}^q \left[\alpha_1(u, u_1)^2 \right]$$

because of $\alpha_1(u_2, u)^2 \alpha_1(u, u_1)^2$ being null as a consequence of having disjoint supports. On the other hand, and bearing in mind the computations we have made to obtain the covariance function for the process $R_n^{lb}(x)$, we have that

$$\mathbf{E}^q \left[\alpha_1\big(u'', u'\big)^2 \right] = \int_{u'}^{u''} v^q(u) \, du,$$

and

$$C \le 3 \int_{u}^{u_2} v^q(u) \, du \int_{u_1}^{u} v^q(u) \, du \le 3 \Big(\int_{u_1}^{u_2} v^q(u) \Big)$$

for $v^q(u) = v^{lb}\Big(F^{lb^{-1}}(u)\Big)$. So taking $H(u)$ as $\int_0^u v^{lb}(u)\,du$ we have the desired bound in equation (21) and hence the process $Q_n^{lb}(x)$ is tight in $D[0,1]$, therefore $R_n^{lb}(x)$ is tight in $D[-\infty,\infty]$. $\diamond$

Proof of Proposition 2.5: First notice that

$$\mathbf{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}\mathbf{g}(x_i)\mathbf{1}_{\{x_i\leq x\}} = \int \frac{1}{y}\mathbf{g}(z)\mathbf{1}_{\{z\leq x\}}dF_n^{lb}(z,y),$$

and for every entry $g_j(x)$ in $\mathbf{g}(x)$, the integral can be written in terms of the empirical process theory as $F^{lb}\,f$, for $f\in\mathcal{C}_j$, that is to say: as an empirical process indexed by the functions in class $\mathcal{C}$ defined by

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

Because of the hypothesis, the functions $f(x)$ are F^{lb}–measurable, and as they are the product of a fixed function $y^{-1}g_j(z)$ and the indicators of $(-\infty,x]$ for $x\in\mathbf{R}$, Lemma 2.6.18 in van der Vaart and Wellner (1996) shows that $\mathcal{C}_j$ is a VC–subgraph class of functions whose envelope verifies

$$e_{\mathcal{C}_j}(x)\leq\frac{1}{y}|g_j(z)|,$$

and hence it is a uniformly bounded function.

As we can find real numbers A, B such that the function $T(f) = Af+B$ takes values into $[0,1]$ for every $f\in\mathcal{C}_j$, the classes of functions $\mathcal{C}_j' = \{T(f)\,:\,f\in\mathcal{C}_j\}$ are also F^{lb}–square–integrable classes of functions, and Theorem 2.14.9 in van der Vaart and Wellner (1996) shows that if $\mathbb{G}_n^F(f)$ for $f\in\mathcal{F}$ denotes the empirical process $\sqrt{n}(\,F_n\,f\,-\,F\,f\,)$ indexed by the class of function $\mathcal{F}$

$$\mathbf{P}\bigg(\Big\|\mathbb{G}_n^{F^{lb}}(f)\Big\|_{\mathcal{C}_j} > Ct\bigg) = \mathbf{P}\bigg(\Big\|\mathbb{G}_n^{F^{lb}}(f)\Big\|_{\mathcal{C}_j'} > C't\bigg) \leq \left(\frac{Dt}{\sqrt{V}}\right)^V e^{-2t^2}.$$

In particular, in the case of $t=\sqrt{\log n}$ we have that

$$\mathbf{P}\bigg(\Big\|\mathbb{G}_n^{F^{lb}}(f)\Big\|_{\mathcal{C}_j} > C\sqrt{\log n}\bigg) \leq C''(\log n)^{V/2}e^{-2\log n} = O\bigg(\frac{1}{n^{2-V\alpha/2}}\bigg)$$

for α such that $0 < V\alpha < 2$, and as a consequence

$$\sum_{n\geq 0}\mathbf{P}\bigg(\Big\|\mathbb{G}_n^{F^{lb}}(f)\Big\|_{\mathcal{C}_j} > C\sqrt{\log n}\bigg) < \infty.$$

Therefore, the Borel–Cantelly Lemma leads to the following result

$$\sup_{x\in\mathbf{R}}\left|\mathbf{G}_n(x)-\frac{1}{\mu_Y}\mathbf{G}(x)\right| = O\left(\sqrt{\frac{\log n}{n}}\right)$$

with probability one, and hence

$$\mathbf{G}_n(x)=\frac{1}{\mu_Y}\mathbf{G}(x)+O\left(\sqrt{\frac{\log n}{n}}\right)$$

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

Now, the result follows from Proposition 2.1 as we have:

$$\begin{aligned}
R_n^{lb^1}(x) &= R_n^{lb}(x)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{y_i}\mathbf{g}(x_i)^T\Big(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n\Big)\mathbf{1}_{\{x_i\leq x\}}\\
&= R_n^{lb}(x)+\frac{1}{\sqrt{n}}n\left(\frac{1}{\mu_Y}\mathbf{G}(x)+O\left(\sqrt{\frac{\log n}{n}}\right)\right)^T\Big(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n\Big)\\
&= R_n^{lb}(x)+\frac{1}{\sqrt{n}}n\frac{1}{\mu_Y}\mathbf{G}(x)^T\Big(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n\Big)+O\left(\frac{\sqrt{\log n\log\log n}}{\sqrt{n}}\right)\\
&= R_n^{lb}(x)-\frac{1}{\sqrt{n}}\mathbf{G}(x)^T\mathbf{L}^{-1}\sum_{i=1}^{n}\mathbf{g}(x_i)\frac{\epsilon_i}{y_i}+O\left(\frac{\log n}{\sqrt{n}}\right)
\end{aligned}$$

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

Proof of Theorem 2.1: As we have seen, the process $R_n^{lb^1}(x)$ can be decomposed in the following way

$$R_n^{lb^1}(x)=R_n^{lb}(x)+R_n^{lb^2}(x)+o\left(\frac{\log n}{\sqrt{n}}\right)$$

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

$$R_n^{lb^2}(x)=-\frac{1}{\sqrt{n}}\mathbf{G}(x)^T\mathbf{L}^{-1}\sum_{i=1}^{n}\mathbf{g}(x_i)\frac{\epsilon_i}{y_i}.$$

The multivariate CLT applied to the summation factor in $R_n^{lb^2}(x)$ shows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}(x_i)\frac{\epsilon_i}{y_i}\to N\Big(\mathbf{0},\boldsymbol{\Sigma}^{lb}\Big),$$

hence finite dimensional distributions of $R_n^{lb^2}(x)$ converge to finite dimensional distributions of a gaussian random process whose expectation is null and whose covariance is given by

$$\mathbf{G}(x)^T\mathbf{L}^{-1}\boldsymbol{\Sigma}^{lb}\mathbf{L}^{-1}\mathbf{G}(x).$$

Now, define $Q_n^{lb^2}(u)$ by means of the same quantile transformation we use in the proof of Proposition 2.4, say $R_n^{lb^2}(x) = Q_n^{lb^2}\big(F^{lb}(x)\big)$. It is obvious that $Q_n^{lb^2}(u)$ is a linear combination of random variables whose behavior is independent from u, and whose coefficients are functions of u, hence $Q_n^{lb^2}(u)$ is tight in the space $C[0,1]$ of continuous functions in $[0,1]$ with the sup norm, and therefore $Q_n^{lb^2}(u)$ is also tight in $D[0,1]$, and $R_n^{lb^2}(x)$ in $D[-\infty,\infty]$.

As we have seen that $R_n^{lb}(x)$ is tight in $D[-\infty,\infty]$ we have that $R_n^{lb^1}(x)$ is tight in $D[-\infty,\infty]$. ◇

Proof of Proposition 2.6: Following the same argumentation that was given in Proposition 2.1 we obtain that

$$\hat{\boldsymbol{\beta}}_n^* = \hat{\boldsymbol{\beta}}_n + \mu_Y \mathbf{L}^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{g}(x_i)\frac{\hat{\epsilon}_i^*}{y_i} + O\left(\frac{\log\log n}{n}\right).$$

Notice that $\hat{\epsilon}_i^* = \hat{\epsilon}_i\gamma_i$ and that

$$\hat{\epsilon}_i = \epsilon_i + \mathbf{g}(x_i)^T\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n\right),$$

and hence

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

Because of the properties of γ_i and $\mathbf{g}(x_i)\mathbf{g}(x_i)^T$ the Law of the Iterated Logarithm can be used to show

$$\frac{1}{n}\sum_{i=1}^n \mathbf{g}(x_i)\mathbf{g}(x_i)^T\frac{\gamma_i}{y_i} = O\left(\sqrt{\frac{\log\log n}{n}}\right).$$

In addition, from Proposition 2.1 we found that $\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n = O\left(\sqrt{\log\log n/n}\right)$, therefore the thesis follows. ◇

Proof of Proposition 2.7: . From equation (14) we see that

$$\begin{aligned} R_n^{lb^1*}(x) &= R_n^{lb*}(x) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{y_i}\mathbf{g}(x_i)^T\left(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^*\right)\mathbf{1}_{\{x_i\le x\}} \\ &= R_n^{lb*}(x) + T_1(x). \end{aligned}$$

By means of the same ideas we use in the proof of Proposition 2.5 and using in this case Proposition 2.6 we can see that

$$T_1(x) = -\mathbf{G}(x)^T\mathbf{L}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{g}(x_i)\frac{\epsilon_i}{y_i}\gamma_i + O\left(\frac{\log n}{\sqrt{n}}\right)$$

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

Again using the expression we gave for $\hat{\epsilon}_i$ in Proposition 2.6 we obtain

$$\begin{aligned} R_n^{lb^*}(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{\epsilon}_i^*}{y_i} \mathbf{1}_{\{x_i \le x\}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\epsilon_i}{y_i} \gamma_i \mathbf{1}_{\{x_i \le x\}} \\ &+ \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{g}(x_i)^T \frac{\gamma_i}{y_i} \mathbf{1}_{\{x_i \le x\}} \right) \left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n \right) = T_2(x) + T_3(x). \end{aligned}$$

First term of $T_3(x)$ in previous expression can be written as an empirical process with respect to the joint empirical distribution $F_n^{lb^*}(z, y, \gamma)$ of the bootstrap and the population samples in the following way

$$\mathbf{G}'_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\gamma_i}{y_i} \mathbf{g}(x_i)^T \mathbf{1}_{\{x_i \le x\}} = \int \frac{\gamma}{y} \mathbf{g}(z)^T \mathbf{1}_{\{z \le x\}} dF_n^{lb^*}(z, y, \gamma).$$

Therefore, the entries in $\mathbf{G}'_n(x)$ are given by $F^{lb^*} f$, an empirical process indexed by functions in the class

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

and, having in mind that Γ (the wild bootstrap r.v.) and $\left(X^{lb}, Y^{lb}\right)$ are independent, and hence $dF^{lb^*}(z, y, \gamma)$ is just $dF^{lb}(z, y) p_\gamma$, besides the properties of Γ it is not difficult to see that this class of functions is a F^{lb^*}–measurable VC–subgraph class of functions in the same manner as shown in Proposition 2.5. Moreover, as Γ has null espectation, we can show that $\mathbf{G}'_n(x) = O\left(\sqrt{\log n}\right)$ with probability one and uniformly for all $x \in \mathbf{R}$, in the same way we did in that case. As $\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n = O\left(\sqrt{\log \log n / n}\right)$ because of Proposition 2.1, we have

$$T_3(x) = O\left(\sqrt{\frac{\log n \log \log n}{n}} \right),$$

from which we obtain the result. ◇

Proof of Theorem 2.2: The result follows from Proposition 2.7 arguing as in Theorem 2.1, taking into account that because of the bootstrap random variable Γ, the random errors ϵ_i become now bootstrap random errors $\epsilon_i^* = \epsilon_i \gamma_i$.

Recall that the distribution of the bootstrap random error $\left(X^{lb}, \varepsilon^{lb} \Gamma\right)$, ε^{lb} being $\left(Y^{lb} - m\left(X^{lb}\right)\right)$, is given by

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

where a and b are the values the wild bootstrap random variable Γ can take and p_a and p_b their respective probabilities. We have also used $F^{lb}(z, e)$ to denote the distribution $\big(X^{lb}, \varepsilon^{lb}\big)$. Therefore, $\big(X^{lb}, \varepsilon^{lb}\Gamma\big)$ has a continuous distribution, with the same first and second order moments $\big(X^{lb}, \varepsilon^{lb}\big)$ has. $\diamond$

5.2 Multivariate case

Proof of Proposition 3.1: From the definition of $R_n^{lb}(\mathbf{x})$ it is clear that this process belongs to $D\big(\mathbf{R}^d\big)$ because any intersection between quadrants(see definition in Bickel and Wichura (1971)) on $\mathbf{R}^d$ and $\{\mathbf{x} \in \mathbf{R}^d : \mathbf{x} \leq \mathbf{x}_i\}$ is again a quadrant in $\mathbf{R}^d$ and the fact that $R_n^{lb}(\mathbf{x})$ is a finite linear combination of the indicators of $\{\mathbf{x} \in \mathbf{R}^d : \mathbf{x} \leq \mathbf{x}_i\}$ whose coefficients are continuous functions in $\mathbf{R}^d$.

The finite dimensional distribution of a vector $\big(R_n^{lb}\big(\mathbf{x}^1\big), \ldots, R_n^{lb}\big(\mathbf{x}^k\big)\big)$ 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 $R_n^{lb}(\mathbf{x})$ is null and again its covariance, as in the case of Proposition 2.4, is given by $\mathbf{E}^w\left[v^{lb}(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i \leq \mathbf{x} \wedge \mathbf{x}'\}}\right]$.

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

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

for

$$Q_n^{lb}(\mathbf{u}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{y_i}\big(y_i - m\big(T^{-1}(\mathbf{u}_i)\big)\big)\mathbf{1}_{\{\mathbf{u}_i \leq \mathbf{u}\}},$$

and $\mathbf{u}_i = T(\mathbf{x}_i)$ for T defined as

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

where we have used $F^{lb}(x_i|x_{i+1}, \ldots, x_d)$ to denote the conditional distribution of the random variable $X_i^{lb}|X_{i+1}^{lb}, \ldots, X_d^{lb}$ and $F^{lb}(x_d)$ to denote 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 Bickel and Wichura (1971), for a quadrant $D = [a_1, a_1 + b_1] \times \cdots \times [a_d, a_d + b_d]$ in $\mathbf{R}^d$ and a function H from $\mathbf{R}^d$ into $\mathbf{R}$ the increment of H around D is defined as

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

In particular, notice that if $H(\mathbf{x}) = \mathbf{1}_{\{\mathbf{x}_j \leq \mathbf{x}\}}$

$$\begin{aligned}
H(D) &= \sum_{l_1=0}^{1} \cdots \sum_{l_d=0}^{1} (-1)^{d-\sum_j l_j} \mathbf{1}_{\{x_1 \leq a_1 + l_1 b_1, \ldots, x_d \leq a_d + l_d b_d\}} \\
&= \big(\mathbf{1}_{\{x_1 \leq a_1 + b_1\}} - \mathbf{1}_{\{x_1 \leq a_1\}}\big) \\
&\qquad \sum_{l_2=0}^{1} \cdots \sum_{l_d=0}^{1} (-1)^{d-\sum_{j=2}^{d} l_j} \mathbf{1}_{\{x_2 \leq a_2 + l_2 b_2, \ldots, x_d \leq a_d + l_d b_d\}} \\
&= \ldots \\
&= \big(\mathbf{1}_{\{x_1 \leq a_1 + b_1\}} - \mathbf{1}_{\{x_1 \leq a_1\}}\big) \;\ldots\; \big(\mathbf{1}_{\{x_d \leq a_d + b_d\}} - \mathbf{1}_{\{x_d \leq a_d\}}\big) \\
&= \mathbf{1}_{\{\mathbf{x}_j \in D\}}
\end{aligned}$$

and, as a consequence of being the process $Q_n^{lb}(\mathbf{u})$ a linear combination of indicators $\mathbf{1}_{\{\mathbf{u}_i \leq \mathbf{u}\}}$, we obtain that for a quadrant $D \subset [0,1]^d$

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

Moreover, for quadrants D_1 and D_2 in $\mathbf{R}^d$ that are neighbouring blocks in $[0,1]^d$(see definition in Bickel and Wichura (1971)):

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

and, using Lemma 5.1 in Stute (1997) with $\alpha_i = \alpha_i(D_1)$ and $\beta_i = \alpha_i(D_2)$ we obtain

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

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

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

and therefore

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

From which we have that

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

where we have taken $\mu(D)$ to be $\sqrt{3}\mathbf{E}^q\left[\alpha_i(D)^2\right]$. Hence, condition (3) in Bickel and Wichura (1971) becomes fulfilled for neighbouring blocks D_1 and D_2 with $\gamma_1 = \gamma_2 = 2$ and $\beta_1 = \beta_2 = 1$, and therefore Q_n^{lb} is tight in $[0,1]^d$ which means that R_n^{lb} 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^{lb}

$$\begin{aligned}
\mu(D) &= \mathbf{E}^q\left[\left(\frac{y_i - m(T^{-1}(\mathbf{u}_i))}{y_i}\right)^2 \mathbf{1}_{\{\mathbf{u}_i \in D\}}\right] \\
&= \mathbf{E}^{lb}\left[\left(\frac{y_i - m(\mathbf{x}_i)}{y_i}\right)^2 \mathbf{1}_{\{\mathbf{x}_i \in T^{-1}(D)\}}\right] \\
&= \int_{T^{-1}(D)} v^{lb}(\mathbf{z})\, dF^{lb}(\mathbf{z}).
\end{aligned}$$

◇

Proof of Theorem 3.1: Follow the same argumentation given in the proof of Theorem 2.1. ◇

Proof of Theorem 3.2: In the multivariate case, the distribution of the bootstrap random error $\left(\mathbf{X}^{lb}, \varepsilon^{lb}\Gamma\right)$, ε^{lb} being $\left(Y^{lb} - m(\mathbf{X}^{lb})\right)$, is given by

$$\mathbf{P}\Big(\mathbf{X}^{lb} \le \mathbf{z}, \varepsilon^{lb}\Gamma \le e\Big) = F^{lb}\Big(\mathbf{z}, \frac{e}{a}\Big)\, p_a + F^{lb}\Big(\mathbf{z}, \frac{e}{b}\Big)\, p_b,$$

where a and b are the values the wild bootstrap random variable Γ and p_a and p_b their respective probabilities as in the proof of Theorem 2.2. Hence the same argumentation given there also works in this case. ◇

5.3 Selection Bias

Proof of Proposition 3.2: Follow the proof of Proposition 2.2, taking into account that

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

◇

Proof of Proposition 3.3: $I^w(\mathbf{x}, y)\overline{w}^H$ is an empirical process indexed by the following class of functions:

$$\mathcal{C} = \{f_{\mathbf{x}}(\mathbf{z}, y) \,:\, f_{\mathbf{x}}(\mathbf{z}, y) = \frac{1}{w(\mathbf{z}, y)}\mathbf{1}_{\{\mathbf{z} \le \mathbf{x}\}}, \quad \mathbf{x} \in \mathbf{R}\}.$$

But as $w(\mathbf{z}, y)^{-1}$ acts as a fixed functions and indicators of quadrants in $\mathbf{R}^d$ are also VC–classes of functions, the proof follows the same argument that was given in the Proposition 2.3 $\diamond$

Proof of Theorem 3.3: The argument is analogous to the one given for the proofs of Proposition 3.1 and Theorem 3.1 but using the reciprocal of w_i instead of the one for y_i. $\diamond$

Proof of Theorem 3.4: Follow the argumentation given for Theorem 2.2 with the changes pointed out in the previous result. $\diamond$

References

Ansell, J. I. and Phillips, M. J. (1994). *Practical methods for reliability data analysis*, volume 14 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York. , Oxford Science Publications.

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.*, **42**, pp. 1656–1670.

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

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

Cristóbal, J. A. and Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. *J. Statist. Plann. Inference*, **89**, pp. 145–168.

Cristóbal, J. A. and Alcalá, J. T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data. *Test*, **10**(2), pp. 309–332.

Cristóbal, J. A., Ojeda, J. L., and Alcalá, J. T. (2004). Confidence bands in nonparametric regression with length biased data. *Ann. Inst. Statist. Math.*, **56**(3), pp. 475–496.

Domínguez, M. A. and Lobato, I. N. (2003). Testing the martingale difference hypothesis. *Econometric Rev.*, **22**(4), pp. 351–377.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**(4), pp. 1926–1947.

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

van Keilegom, I., Sánchez Sellero, C., and Gonzlez-Manteiga, W. (2007). Goodness–of–fit test in parametric regression based on the estimation of the error distribution. *TEST*.

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.*, **16**(4), pp. 1696–1708.

Navarro, J., Ruiz, J. M., and del Aguila, Y. (2001). Parametric estimation from weighted samples. *Biom. J.*, **43**(3), pp. 297–311.

Ojeda, J. L., Cristóbal, J. A., and Alcalá, J. T. (2004). Nonparametric confidence bands construction for GLM models with length biased data. *J. Nonparametr. Stat.*, **16**(3-4), pp. 421–441.

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

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

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**(2), pp. 179–189.

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

Quesenberry, Jr., C. P. and Jewell, N. P. (1986). Regression analysis based on stratified samples. *Biometrika*, **73**(3), pp. 605–614.

Rao, C. R. (1997). *Statistic and True. Putting chance to work*. World Scientific Publishing, 2nd. edition.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.*, **25**(2), pp. 613–641.

Stute, W., González Manteiga, W., and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.*, **93**(441), pp. 141–149.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer-Verlag, New York. With applications to statistics.

Wu, C. O. (2000). Local polynomial regression with selection biased data. *Statist. Sinica*, **10**(3), pp. 789–817.

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

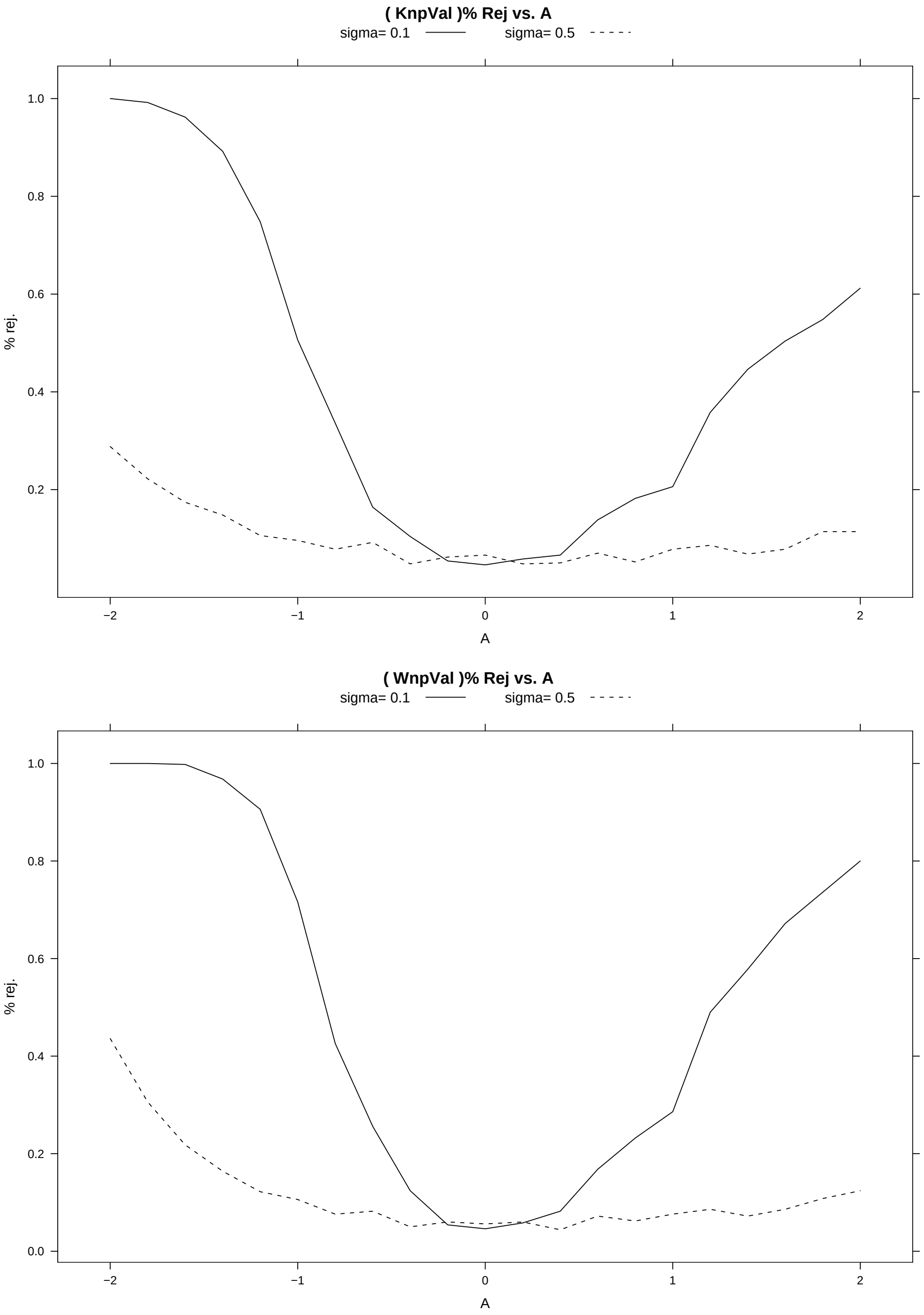

Figure 2: Power function for example 4.2($\alpha = 0.05$).

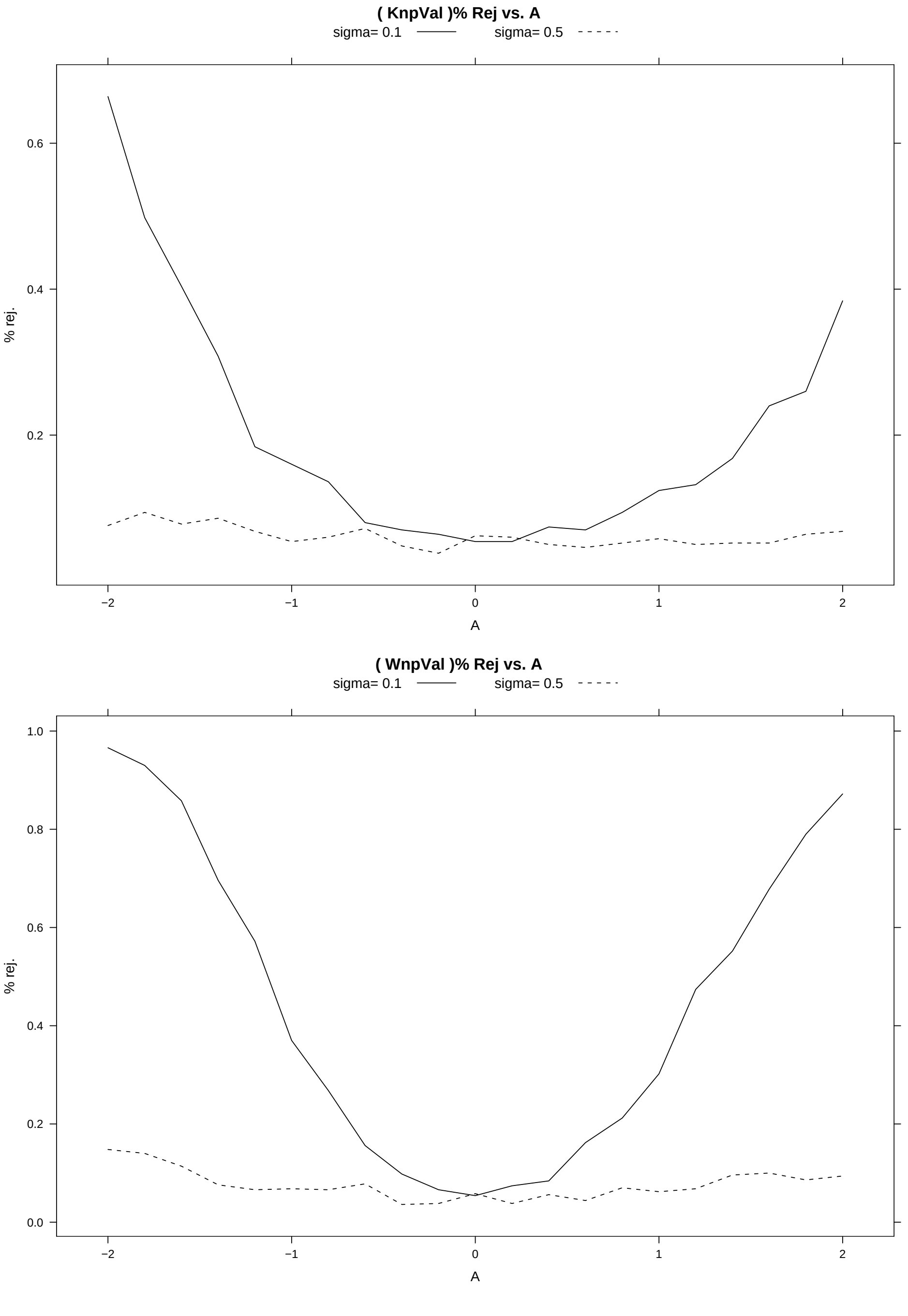

Figure 3: Power function for example 4.3($\alpha = 0.05$).

Reports in Statistics and Operations Research

2004

04-01 Goodness of fit test for linear regression models with missing response data. *González Manteiga, W., Pérez González, A.*
Canadian Journal of Statistics (to appear).

04-02 Boosting for Real and Functional Samples. An Application to an Environmental Problem. *B. M. Fernández de Castro and W. González Manteiga*.

04-03 Nonparametric classification of time series: Application to the bank share prices in Spanish stock market. *Juan M. Vilar, José A. Vilar and Sonia Pértega*.

04-04 Boosting and Neural Networks for Prediction of Heteroskedatic Time Series. *J. M. Matías, M. Febrero, W. González Manteiga and J. C. Reboredo.*

04-05 Partially Linear Regression Models with Farima-Garch Errors. An Application to the Forward Exchange Market. *G. Aneiros Pérez, W. González Manteiga and J. C. Reboredo Nogueira.*

04-06 A Flexible Method to Measure Synchrony in Neuronal Firing. *C. Faes, H. Geys, G. Molenberghs, M. Aerts, C. Cadarso-Suárez, C. Acuña and M. Cano.*

04-07 Testing for factor-by-curve interactions in generalized additive models: an application to neuronal activity in the prefrontal cortex during a discrimination task. *J. Roca-Pardiñas, C. Cadarso-Suárez, V. Nacher and C. Acuña.*

04-08 Bootstrap Estimation of the Mean Squared Error of an EBLUP in Mixed Linear Models for Small Areas. *W. González Manteiga, M. J. Lombardía, I. Molina, D. Morales and L. Santamaría.*

04-09 Set estimation under convexity type assumptions. *A. Rodríguez Casal.*

2005

05-01 SiZer Map for Evaluating a Bootstrap Local Bandwidth Selector in Nonparametric Additive Models. *M. D. Martínez-Miranda, R. Raya-Miranda, W. González-Manteiga and A. González-Carmona.*

05-02 The Role of Commitment in Repeated Games. *I. García Jurado, Julio González Díaz*.

05-03 Project Games. *A. Estévez Fernández, P. Borm, H. Hamers*

05-04 Semiparametric Inference in Generalized Mixed Effects Models. *M. J. Lombardía, S. Sperlich*

2006

06-01 A unifying model for contests: effort-prize games. J. González Díaz

06-02 The Harsanyi paradox and the "right to talk" in bargaining among coalitions. J. J. Vidal Puga

06-03 A functional analysis of NOx levels: location and scale estimation and outlier detection. M. Febrero, P. Galeano, W. González-Manteiga

06-04 Comparing spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga

06-05 On the spectral simulation of spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal

06-06 An L_2-test for comparing spatial spectral densities. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

2007

07-01 Goodness-of-fit tests for the spatial spectral density. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

07-02 Presmothed estimation with left truncated and right censores data. M. A. Jácome, M. C. Iglesias-Pérez

07-03 Robust nonparametric estimation with missing data. G. Boente, W. González-Manteiga, A. Pérez-González

07-04 k-Sample test based on the common area of kernel density estimators, P. Martínez-Camblor, J. de Uña Álvarez, N. Corral-Blanco

07-05 A bootstrap based model checking for selection-biased data, J. L. Ojeda, W. González-Manteiga, J . A. Cristobal

Previous issues (2001 – 2003):
http://eio.usc.es/pub/reports.html