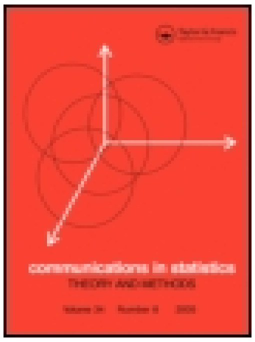

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

High breakdown point robust estimators with missing data

Florencia Statti, Mariela Sued & Victor J. Yohai

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2017, VOL. 0, NO. 0, 1–18
https://doi.org/10.1080/03610926.2017.1388396

High breakdown point robust estimators with missing data

Florencia Statti[a], Mariela Sued[a], and Victor J. Yohai[b]

[a]Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, CONICET, Argentina; [b]Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina

ABSTRACT

In this paper, we propose a new procedure to estimate the distribution of a variable y when there are missing data. To compensate the presence of missing responses, it is assumed that a covariate vector $\mathbf{x}$ is observed and that y and $\mathbf{x}$ are related by means of a semi-parametric regression model. Observed residuals are combined with predicted values to estimate the missing response distribution. Once the responses distribution is consistently estimated, we can estimate any parameter defined through a continuous functional T using a plug in procedure. We prove that the proposed estimators have high breakdown point.

ARTICLE HISTORY
Received 25 May 2016
Accepted 27 September 2017

KEYWORDS
Asymptotic distribution; breakdown point; location and dispersion functionals; missing at random; quantiles.

MATHEMATICS SUBJECT CLASSIFICATION
62F35

1. Introduction

Missing data arise in many situations across different fields, from natural sciences to humanities. Survey methods have already faced this problem from their very beginning, and considerable efforts have been invested since the early 1970s up to these days to develop data analysis procedures for missing data.

Another problem that should be faced is the presence of outliers in the sample. Many of the most popular methods, as maximum likelihood estimators, are very sensitive to the occurrence of atypical observations. Estimators which are not much influenced by outliers are called robust estimators. The main purpose of this article is to provide robust estimators that can be used when there are missing values in the sample.

The estimation of the mean of a given random variable y, based on a sample with missing observations, has acquired a remarkable place in modern statistics. To deal with this problem it is often assumed that for each subject i in the sample, a vector $\mathbf{x}_i$ of p covariates is always observed, even in the case that the corresponding scalar response y_i is missing. The missing at random (MAR) assumption (Rubin 1976) allows to identify the parameter of interest in terms of the distribution of the observed data. It states that the distribution of the missing mechanism does not depend on the variable of interest, once the covariates are available. To be more precise, let a_i be the indicator of whether y_i is observed. Namely, $a_i = 1$ whenever the response y_i is observed in subject i. The MAR hypothesis establishes that

$$\mathrm{P}(a = 1|\mathbf{x}, y) = \mathrm{P}(a = 1|\mathbf{x}) = \pi(\mathbf{x}) \tag{1}$$

CONTACT Mariela Sued marielasued@gmail.com Intendente Guiraldes 2160, Ciudad Universitaria, Pabellón II, 2do. piso (C1428EGA), Buenos Aires, Argentina.

Supplemental data for this article can be accessed on the publisher's website at https://doi.org/10.1080/03610926.2017.1388396

where $(\mathbf{x}^{\mathrm{t}}, a, y)$ is a random vector distributed as $(\mathbf{x}_i^{\mathrm{t}}, a_i, y_i)$. Condition (1) implies that y is independent of a given $\mathbf{x}$ and therefore, $\mathrm{E}[y] = \mathrm{E}[\mathrm{E}[y|\mathbf{x}, a = 1]]$. This representation of $\mathrm{E}[y]$ suggests to postulate a parametric model $g(\mathbf{x}, \boldsymbol{\beta}_0)$ for the regression function $\mathrm{E}[y|\mathbf{x}, a = 1]$ and estimate $\mathrm{E}[y]$ by the so-called outcome regression estimator (OR), defined as the mean value of the predicted responses under the model: $n^{-1}\sum g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_n)$, where $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}_0$ based on those individuals with $a_i = 1$.

There exist other methods for estimating the mean of y in presence of missing data. For instance, we can mentioned the inverse probability weighted (IPW) procedures which are based on the estimation of the propensity $\pi(\mathbf{x})$, defined in Equation (1). There exist also the so-called doubly protected estimators that remain consistent when either (but not necessarily both) the considered model for the propensity score $\mathrm{P}(a = 1|\mathbf{x})$ or the model postulated for the regression $\mathrm{E}[y|\mathbf{x}, a = 1]$ is correctly specified. See, for instance, Bang and Robins (2005) and Rotnitzky et al. (2012). These estimators seems to be difficult to robustify and for this reason we will focus on outcome regression estimators.

Causal inference has necessary to deal with missing data. The potential outcomes model, popularized by Rubin (1974), can be used to quantify the effect of two different treatments, say t_0 and t_1, on some response of interest. In this model, two potential outcomes (or counterfactual variables) $y^{(0)}$ and $y^{(1)}$ are defined as the response of a participant in the case she/he were exposed to t_0 or t_1, respectively. The average treatment effect is defined as $\mathrm{E}[y^{(1)}] - \mathrm{E}[y^{(0)}]$ and so, the estimation of each mean $\mathrm{E}[y^{(j)}]$, $j = 0, 1$ is required. However, we never observe both potential outcomes in the same subject of the sample. We only observe the outcome corresponding to the assigned treatment while the other one (counterfactual) remains missing. In this way, the estimation of $\mathrm{E}[y^{(j)}]$, $j = 0, 1$ requires to use missing data techniques, since $y^{(j)}$ is only available for those individuals with treatment level $T = t_j$, and is missing for those with $T \neq t_j$. Westreich et al. (2015) present an interesting discussion on the connections between causal inference and missing data.

It is well known that, even when all observations are available, the mean is very sensitive to the presence of outliers in the sample. Just one outlying observation can take this estimator beyond any limit. Robust procedures have been developed to overcome this limitation. For instance, the median, which is probably one of the most famous robust location parameter, does not suffer the instability phenomenon described for the mean. We can consider many other robust location estimators. We can mentioned, for examples, M-, L- and R -estimators which have different degrees of robustness. This suggests that instead of estimating the mean we can attempt to estimate other location parameters defined as $\mu_0 = T(F_0)$, where T is a continuous location functional and F_0 is the distribution of y. Several authors have deal with this problem in presence of missing data. Bianco et al. (2010) obtained consistent estimators of M-location functionals assuming a partially linear model to describe the relationship between y and $\mathbf{x}$. This approach requires that both the distribution of the response y and that of the regression error under the true model, to be symmetric.

Going back to the causal inference framework, Zhang et al. (2012), focused on the estimation of the median or, more generally, on the estimation of any quantile associated with the distribution of counterfactual variables. They presented an outcome regression estimator using a parametric model for the conditional distribution of y given $\mathbf{x}$ and an IPW estimator, which assumes a parametric model for the propensity score $\pi(\mathbf{x})$, defined in Equation (1). They have also considered a doubly protected estimator.

Sued and Yohai (2013), SY in the remainder of the paper, proposed a consistent estimator of F_0 under a semi-parametric regression model $y = g(\mathbf{x}, \boldsymbol{\beta}_0) + u$, where the distribution of the error term u is completely unspecified. This approach allows for the consistent estimation of any parameter defined by a weakly continuous functional at F_0. Particular emphasis is dedicated in SY to location functionals such as the median (and any quantile), α-trimmed means (and any L-location functional) as well as M-location functionals. The consistency of these procedures neither requires the symmetry assumptions used by Bianco et al. (2010), nor the parametric model postulated by Zhang et al. (2012) for constructing outcome regression estimators.

In this paper we propose a new estimator $\widehat{F_n}$ of F_0. Once consistently estimated F_0, we can robustly estimate any parameter $\theta_0 = T(F_0)$, where T is a functional weakly continuous at F_0, by means of a plug in procedure. Note that, from now on, μ_0 indicates location parameters while θ_0 denotes a generic one. We prove that the estimator $\widehat{\theta_n} = T(\widehat{F_n})$ of θ_0 has a larger breakdown point than that of the estimator $T(\widetilde{F_n})$, where $\widetilde{F_n}$ is the estimator of F_0 proposed in SY. Breakdown point (BP) is an important measure of robustness. Roughly speaking, *the BP is the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values*, as mentioned in Donoho and Huber (1983).

High BP estimates are desirable, even if the proportion of outliers is not expected to be large, because they typically over perform the behavior of estimators with smaller BP. For instance, the median which has a BP equal to 0.5, is the estimator with the least asymptotic bias for any level of contamination, see Huber (1964).

When all observations y_i, $1 \le i \le n$ are available, θ_0 is typically estimated with $T(F_n)$, where F_n is the empirical distribution of y_i. A remarkable property of the estimator presented in this work is that its BP can be as high as that of $T(F_n)$ when the fraction of missing data is close to 0. In the later case, for instance, the BP for estimating the median converges to 0.5. On the other hand, when the fraction of missing data goes to 1 the BP converges to that obtained in SY, which does not depend on the fraction of missing responses.

This paper is organized as follows. In Sec. 2, we introduce the new estimator of the distribution of y which is used to estimate any parameter defined by means of a continuous functional through a plug in procedure. We establish the consistency and the asymptotic normality of the proposed procedure for regular functionals. We also consider estimators of the quantiles of the distribution of y and establish their asymptotic properties. In Sec. 4, we obtain a lower bound for the breakdown point of the estimators presented here. In Sec. 5.1, we report the results of a Monte Carlo simulation both for location and dispersion parameters. This last problem has not been considered by any of the aforementioned works. In Sec. 5.2, the result of a real data study are presented. In Sec. 6, we compare the asymptotic variance of the new procedure with that of SY. Some conclusions are drawn in Sec. 7. The Supplementary Material (SM), available on line, contains all the proofs, and the tables and figures corresponding to Sec. 5. A code for the computation of the estimators presented here and in SY, used in the Monte Carlo Study, is also included in the SM.

2. Estimating the distribution of *y*

Let $(\mathbf{x}_i^{\mathrm{t}}, a_i, y_i)$, $1 \le i \le n$, be independent identically distributed (i.i.d) vectors, and consider $(\mathbf{x}^{\mathrm{t}}, a, y)$ with the same distribution as $(\mathbf{x}_i^{\mathrm{t}}, a_i, y_i)$. Recall that F_0 denotes the marginal distribution of the outcome y. Let T be a weakly continuous functional at F_0 and suppose that we are interested in estimating the parameter $\theta_0 = T(F_0)$. As it was already mentioned, when all the responses y_i, $1 \le i \le n$, are available, θ_0 can be estimated by $T(F_n)$, where F_n is the empirical distribution given by $F_n = n^{-1} \sum_{i=1}^{n} \delta_{y_i}$, while δ_w denotes the distribution function of a mass point concentrated at w. When some responses are missing, the empirical distribution of the observed responses,

$$\widehat{F}_{0,1} = \frac{1}{\sum_{j=1}^{n} a_j} \sum_{j=1}^{n} a_j \delta_{y_j} \tag{2}$$

converges to $F_{0,1}$, the conditional distribution of y given $a = 1$. In general, $F_{0,1}$ is different from the distribution of y, except when a is independent of y (missing completely at random). To overcome this problem, $\widehat{F}_{0,1}$ should be combined with a consistent estimator of $F_{0,0}$, the conditional distribution of y given $a = 0$. At this point, let us assume a a semi-parametric regression model similar to the one consider in the Introduction for the OR estimator of $\mathrm{E}[y]$. That is, we assume that

$$y = g(\mathbf{x}, \boldsymbol{\beta}_0) + u \tag{3}$$

with $y, u \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^p$, u independent of $\mathbf{x}$, $\boldsymbol{\beta}_0 \in B \subset \mathbb{R}^q$, $g : \mathbb{R}^p \times B \to \mathbb{R}$. We further assume that u is independent of $(\mathbf{x}, a)$. Note that this last assumption implies the MAR condition. All the presented assumptions imply that $F_{0,0} = R_0 * K_0$, where R_0 and K_0 denote the conditional distribution of $g(\mathbf{x}, \boldsymbol{\beta}_0)$ given $a = 0$ and the distribution of u, respectively. Then, we can estimate $F_{0,0}$ by $\widehat{F}_{0,0} = \widehat{R}_n * \widehat{K}_n$, where $\widehat{R}_n$ and $\widehat{K}_n$ are consistent estimators of R_0 and K_0, respectively, while $*$ denotes the convolution of distributions. Let $\eta_n = \sum_{j=1}^{n} a_j/n$, then we can estimate F_0 by

$$\widehat{F}_n = \eta_n \widehat{F}_{0,1} + (1 - \eta_n)\widehat{R}_n * \widehat{K}_n \tag{4}$$

R_0 and K_0 have to be estimated so that $T(\widehat{F}_n)$ turns out to be a robust estimator of θ_0. For this purpose, we need a robust strongly consistent estimator of $\boldsymbol{\beta}_0$, which will be denoted by $\widehat{\boldsymbol{\beta}}_n$. This estimator may be, for example, an S-estimator (see Rousseeuw and Yohai 1984) or an MM-estimator. MM-estimators where introduced for linear models by Yohai (1987) and Fasano et al. (2012) extended them for non linear regression. Since u is independent of a, $\widehat{\boldsymbol{\beta}}_n$ may be obtained by a robust fit of model (3) using the observations $(\mathbf{x}_i, y_i)$ with $a_i = 1$. Let $A = \{i : a_i = 1\}$ and $m = \#A$. Consider the residuals $\widehat{u}_i = y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_n)$, $i \in A$. We will show that we can take $\widehat{R}_n$ as the empirical distribution of $g(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_n)$, $j \notin A$, and $\widehat{K}_n$ as the empirical distribution of $\widehat{u}_i$, $i \in A$. Note that we can write

$$\widehat{R}_n = \frac{1}{n - m}\sum_{j=1}^{n}(1 - a_j)\delta_{g(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_n)}, \quad \widehat{K}_n = \frac{1}{m}\sum_{i=1}^{n} a_i \delta_{\widehat{u}_i}$$

and so $\widehat{R}_n * \widehat{K}_n$ is the empirical distribution of the pseudo-observations $\widehat{y}_{ij} = g(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_n) + \widehat{u}_i$ defined, in principle, for $j \notin A$ and $i \in A$. Then, according to (4), we propose here to estimate F_0 by

$$\widehat{F}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j} a_j + \frac{1}{n\,m}\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{\widehat{y}_{ij}} a_i (1 - a_j) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{y_j} a_i a_j + \delta_{\widehat{y}_{ij}} a_i (1 - a_j)) \tag{5}$$

Namely, $\widehat{F}_n$ assigns mass $1/(nm)$ to each of the nm points $\widehat{y}_{ij}$, $1 \le j \le n$ and $i \in A$, defined by

$$\widehat{y}_{ij} = \begin{cases} y_j & \text{if } j \in A, i \in A \\ g(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_n) + \widehat{u}_i & \text{if } j \notin A, i \in A \end{cases} \tag{6}$$

Finally, we estimate θ_0 by

$$\widehat{\theta}_n = T(\widehat{F}_n) \tag{7}$$

We recall that in SY F_0 is estimated by

$$\widetilde{F}_n = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i \delta_{\widetilde{y}_{ij}} \tag{8}$$

where $\widetilde{y}_{ij} = g(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_n) + \widehat{u}_i,\ 1 \le j \le n, i \in A$. Note that $\widehat{F}_n$ and $\widetilde{F}_n$ are both empirical distributions corresponding to the samples $\mathcal{S}_1 = \{\widehat{y}_{ij}, i \in A, 1 \le j \le n\}$ and $\mathcal{S}_2 = \{\widetilde{y}_{ij}, i \in A, 1 \le j \le n\}$, respectively. Then, if the regression coefficients are estimated using a robust procedure, $\mathcal{S}_2$ contains in general more outliers than $\mathcal{S}_1$. In fact, given j so that $a_j = 1$ and y_j is not an outlier, then none of the values $\widehat{y}_{ij}$, $i \in A$ is an outlier. Instead, for the same j, all the values $\widetilde{y}_{ij}$, where $i \in A$ and $\mathbf{x}_i$ is an outlier, are outliers. This heuristic argument explains why $\widehat{\theta}_n = T(\widehat{F}_n)$ is plausible more robust than $\widetilde{\theta}_n = T(\widetilde{F}_n)$. A formal statement of this fact can be found in Lemma 1.

Observe also that $\widehat{F}_{0,1}$ does not depend on the regression model (3), while both estimators of $F_{0,0}$ and $F_{0,1}$ in SY depend on g. As a consequence, the new estimator is less sensitive to misspecification of the function $g(\mathbf{x}, \boldsymbol{\beta})$ chosen for the regression model (3).

To finish this section we want to emphasize on the semi-parametric nature of the regression model (3), where no parametric family is postulated for the distribution of the error term u.

3. Asymptotic properties

Sued and Yohai (2013) proved that to identify $\boldsymbol{\beta}_0$, it is sufficient to assume that

$$\mathrm{P}\left(g(\mathbf{x}, \boldsymbol{\beta}_0) = g(\mathbf{x}, \boldsymbol{\beta}) + \alpha | a = 1\right) < 1 \tag{9}$$

for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and for all α. No further conditions on K_0 are imposed: we require neither that (i) K_0 is symmetric around 0 nor that (ii) K_0 satisfies a centering condition (e.g., $\mathrm{E}_{K_0}[u] = 0$). To satisfy condition (9) it is required that, in case there is an intercept, it should be included in the error term u and should not be included as one of the components of β_0. For a linear regression model, where $g(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathrm{t}}\mathbf{x}$, condition (9) means that the distribution of the vector $\mathbf{x}$ given $a = 1$ is not concentrated on any hyperplane.

The following assumptions on the function $g(\mathbf{x}, \boldsymbol{\beta})$, the estimator $\widehat{\boldsymbol{\beta}}_n$ and the functional T, all of them already considered in SY, are required to prove the consistency and asymptotic distribution of the estimator $\widehat{\theta}_n$ defined in (7).

A0. The function $g(\mathbf{x}, \boldsymbol{\beta})$ is continuously differentiable with respect to $\boldsymbol{\beta}$ and there exists $\delta > 0$ such that

$$\mathrm{E}\left[\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le \delta} \left\|\dot{g}(\mathbf{x}_1, \boldsymbol{\beta})\right\|^2\right] < \infty, \tag{10}$$

where $\dot{g}(\mathbf{x}, \boldsymbol{\beta})$ denotes the vector of first derivatives of g with respect to $\boldsymbol{\beta}$.

A1. $\{\widehat{\boldsymbol{\beta}}_n\}$ is strongly consistent for $\boldsymbol{\beta}_0$.

A2. The regression estimator $\widehat{\boldsymbol{\beta}}_n$ satisfies $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^{n} a_i I_R(\mathbf{x}_i, y_i) + o_P(1)$ for some function $I_R(\mathbf{x}, y)$ with $\mathrm{E}[aI_R(\mathbf{x}, y)] = 0$ and finite covariance matrix.

A3. T is weakly continuous at F_0.

A4. The following expansion holds: $\sqrt{n}(T(\widehat{F}_n) - T(F_0)) = \sqrt{n}\,\mathrm{E}_{\widehat{F}_n}[I_{T,F_0}(y)] + o_P(1)$, where I_{T,F_0} is the influence function, see Hampel (1974), of T at F_0. We assume also that

$\mathrm{E}_{F_0}[I_{T,F_0}(y)] = 0$, $\mathrm{E}_{F_0}[I^2_{T,F_0}(y)] < \infty$ and I_{T,F_0} is differentiable with bounded derivative $I'_{T,F_0}(\cdot)$.

Remarks 1 and 2 in the Supplementary Material briefly discuss on the validity of A1–A4. The following theorems establish the consistency and the asymptotic normality of $\widehat{\theta}_n = T(\widehat{F}_n)$, defined in (7).

Theorem 1. *Let $\widehat{F}_n$ be defined as in (5) and assume that A0 and A1 hold. Then (a) $\{\widehat{F}_n\}$ converges weakly to F_0 almost sure (a.s.); that is, $\mathrm{P}(\widehat{F}_n \to_w F_0) = 1$. (b) Assume also that A3 holds; then $\widehat{\theta}_n = T(\widehat{F}_n)$ converges a.s. to $\theta_0 = T(F_0)$.*

Theorem 2. *Assume A0–A4. Then, $n^{1/2}(\widehat{\theta}_n - \theta_0) \to_d N(0, \tau^2)$, where*

$$\tau^2 = \frac{1}{\eta^2}\mathrm{E}[\{e(\mathbf{x}_1, u_1, a_1) + f(\mathbf{x}_1, a_1) + a_1\mathbf{c}' I_R(\mathbf{x}_1, u_1)\}^2], \qquad \eta = \mathrm{E}[a_1] \tag{11}$$

$$\mathbf{c} = \mathrm{E}\left[a_1(1-a_2)I'_{T,F_0}(y_1 - g(\mathbf{x}_1, \boldsymbol{\beta}_0) + g(\mathbf{x}_2, \boldsymbol{\beta}_0))\left\{\dot{g}(\mathbf{x}_2, \boldsymbol{\beta}_0) - \dot{g}(\mathbf{x}_1, \boldsymbol{\beta}_0)\right\}\right]$$

$$e(\mathbf{x}_i, u_i, a_i) = a_i I_{T,F_0}(y_i)\mathrm{E}[a_j] + a_i\mathrm{E}[(1-a_j)I_{T,F_0}(g(\mathbf{x}_j, \boldsymbol{\beta}_0) + u_i)|u_i]$$

$$f(\mathbf{x}_i, a_i) = a_i\mathrm{E}[a_j I_{T,F_0}(y_j)] + (1-a_i)\mathrm{E}[a_j]\mathrm{E}[I_{T,F_0}(g(\mathbf{x}_i, \boldsymbol{\beta}_0) + u_j)|\mathbf{x}_i] \tag{12}$$

In many situations, as discussed by Zhang et al. (2012), the parameter of interest is the median, or more generally, any quantile of the distribution F_0. The p-quantile of a distribution F can be defined through the functional $T_p(F) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$. T_p is continuous at F_0 if, for instance, F_0 has a positive density f_0 in a neighborhood of $T_p(F_0)$. In such a case, we obtain that $\widehat{\mu}_n = T_p(\widehat{F}_n)$ converges to $\mu_0 = T_p(F_0)$, with $\widehat{F}_n$ defined by (5). Nevertheless, the influence function of T_p does not satisfy the regularity assumptions regarding its influence function. When $p = 0.5$, $\mu_0 = T_{0.5}(F_0)$ is the median. SY proposed to estimate it by $T_{0.5}(\widetilde{F}_n)$, for $\widetilde{F}_n$ defined in (8) and give a rigorous proof of the asymptotic behavior of the estimator. We adapt the arguments of this proof to obtain the asymptotic distribution of the new estimator $\widehat{\mu}_n = T_p(\widehat{F}_n)$. Then we can state the following theorem.

Theorem 3. *Suppose that F_0 has a positive density f_0 in a neighborhood of $\mu_0 = T_p(F_0)$. Besides assume that A0–A1 holds. Then (a) $\widehat{\mu}_n = T_p(\widehat{F}_n) \to \mu_0$ a.s. (b) Assume also that A2 holds, that F_0 and K_0 have continuous and bounded densities f_0 and k_0 respectively. Then $n^{1/2}(\widehat{\mu}_n - \mu_0) \to_d N(0, \tau^2)$ where τ^2 is as in Theorem 2, with $\mathbf{c}$ replaced by*

$$\mathbf{c}^* = \frac{\mathrm{E}\left[a_1(1-a_2)k_0(-g(\mathbf{x}_2, \beta_0) + \mu_0)(\dot{g}(\mathbf{x}_2, \beta_0) - \dot{g}(\mathbf{x}_1, \beta_0))\right]}{f_0(\mu_0)} \tag{13}$$

and $I_{T_p,F_0}(y)$ replaced by $I_{T_p,F_0}(y) = -\mathrm{sign}_p(y - \mu_0)f_0^{-1}(\mu_0)$, where $\mathrm{sign}_p(y) = 1 - p$, 0 or $-p$ according to $y < 0$, $y = 0$ or $y > 0$, respectively.

4. Breakdown point

SY extended the notion of *Finite Sample Breakdown Point* (FSBP) of an estimator, introduced by Donoho and Huber (1983), to the case where there are missing observations as follows. Let $\mathbf{W} = \{(\mathbf{x}_1, y_1, a_1), \ldots (\mathbf{x}_n, y_n, a_n)\}$ be the set of all observations and missingness indicators. Recall that $A = \{i : 1 \leq i \leq n,\ a_i = 1\}$ and $m = \#A$. Denote by $\mathcal{W}_{ts}$ the set of all samples obtained from $\mathbf{W}$ where at most t points are replaced by outliers, with at most s of these replacements corresponding to the non missing observations. Then $\mathbf{W}^* =$

$\{(\mathbf{x}_1^*, y_1^*, a_1), \ldots, (\mathbf{x}_n^*, y_n^*, a_n)\}$ belongs to $\mathcal{W}_{ts}$ if

$$\sum_{i\in A} I_{\{(\mathbf{x}_i^*, y_i^*)\neq(\mathbf{x}_i, y_i)\}} + \sum_{i\in A^C} I_{\{\mathbf{x}_i^*\neq\mathbf{x}_i\}} \le t \qquad \text{and} \qquad \sum_{i\in A} I_{\{(\mathbf{x}_i^*, y_i^*)\neq(\mathbf{x}_i, y_i)\}} \le s \tag{14}$$

Given an estimator $\widehat{\theta}_n$ of θ_0, we define

$$M_{ts} = \sup_{\mathbf{W}^*\in\mathcal{W}_{ts}} \left|\widehat{\theta}_n(\mathbf{W}^*)\right| \quad \text{and} \quad \kappa(t,s) = \max\left(\frac{t}{n}, \frac{s}{m}\right)$$

The finite sample breakdown point (FSBP) of an estimator $\widehat{\theta}_n$ at $\mathbf{W}$ is given by $\varepsilon^* = \min\{\kappa(t,s) : M_{ts} = \infty\}$. Then, ε^* is the minimum fraction of outliers in the complete sample or in the set of non missing observations required to take the estimator beyond any bound.

In order to obtain a lower bound for the FSBP of the estimator $\widehat{\theta}_n$ introduced in (7), let us recall the definition of uniform asymptotic breakdown point (UABP) ε_U^* of the functional T, introduced in SY. The UABP is defined as the supremum of all $\varepsilon > 0$ satisfying the following property: for all $M > 0$ there exists $K > 0$ depending on M so that $\mathrm{P}_F(|y| \le M) > 1 - \varepsilon$ implies $|T(F)| < K$.

Lower bounds for the UABP of L- and M-location functionals are presented in SY. In particular, in Theorem 8 it is shown that it is possible to define location M-estimators with UABP equal 0.5. This result can also be extended to the dispersion functional defined in Equation (18). The following theorem gives a lower bound for the FSBP of $\widehat{\theta}_n$, defined in (7).

Theorem 4. *Given $\mathbf{W} = \{(\mathbf{x}_1, y_1, a_1), \ldots.(\mathbf{x}_n, y_n, a_n)\}$, let $\mathbf{Z} = \{(\mathbf{x}_i, y_i) : i \in A\}$. Suppose that $\widehat{\boldsymbol{\beta}}_n = \widetilde{\boldsymbol{\beta}}_m(\mathbf{Z})$, where $\widetilde{\boldsymbol{\beta}}_m$ is a regression estimator for samples of size m. Let $\varepsilon_1 > 0$ be a lower bound of the FSBP at $\mathbf{Z}$ of $\widetilde{\boldsymbol{\beta}}_m$ and let $\varepsilon_2 > 0$ be a lower bound of the UABP of T. Then the FSBP ε^* of the estimator $\widehat{\theta}_n$ at $\mathbf{W}$, satisfies the following inequality:*

$$\varepsilon^* \ge \varepsilon_3^{\mathrm{NEW}} = \min\left(\varepsilon_1, 1 - \frac{\sqrt{\eta_n^2 + 4(1-\eta_n)(1-\varepsilon_2)} - \eta_n}{2(1-\eta_n)}\right) \tag{15}$$

where $\eta_n = \frac{m}{n}$.

Next, we analyze the behavior of $\varepsilon_3^{\mathrm{SY}}$ and $\varepsilon_3^{\mathrm{NEW}}$, the lower bound for the FSBP of $\widetilde{\theta}_n$ and $\widehat{\theta}_n$, respectively. Besides, it is shown the behavior of $\varepsilon_3^{\mathrm{NEW}}$ when the fraction of missing data fraction converges to 0 and 1.To express the following results in terms of the fraction of missing data in the sample, let $\delta_n = 1 - \eta_n$ and let $\varepsilon_3^{\mathrm{SY}}(\delta_n)$ be defined as in (15) with $\eta_n = 1 - \delta_n$

Lemma 1. *Under the assumptions of Theorem 4, we have that*

(a) $\varepsilon_3^{\mathrm{SY}} \le \varepsilon_3^{\mathrm{NEW}}(\delta_n)$*, for all* δ_n*,* ε_1 *and* ε_2*.*
(b) $\lim_{\delta_n\to 1} \varepsilon_3^{\mathrm{NEW}}(\delta_n) = \varepsilon_3^{\mathrm{SY}}$*, for all* ε_1 *and* ε_2*.*
(c) $\lim_{\delta_n\to 0} \varepsilon_3^{\mathrm{NEW}}(\delta_n) = \min(\varepsilon_1, \varepsilon_2)$*, for all* ε_1 *and* ε_2*.*

Remark 1. (a) shows that the new estimator is resistant to a larger fraction of outliers and (b) shows that, choosing a regression estimate with $\varepsilon_1 = 0.5$, as in the case of the MM-estimator used in the simulation study, $\varepsilon_3^{\mathrm{NEW}}$ approaches to ε_2 when the missing fraction tends 0. Observe that ε_2 is the UABP of the functional T, which is the asymptotic breakdown point of $T(F_n)$, where F_n is the empirical distribution of y_i, $1 \le i \le n$. (c) means that when the missing fraction tends to 1, $\varepsilon_3^{\mathrm{NEW}}$ approaches the lower bound presented in (a), which is $\varepsilon_3^{\mathrm{SY}}$. Note that $\varepsilon_3^{\mathrm{SY}}$ does not depend on the fraction of missing data. These facts are illustrated in Figure 1, where we plot $\varepsilon_3^{\mathrm{SY}}$ and $\varepsilon_3^{\mathrm{NEW}}$ as a function of δ_n, fixing $\varepsilon_1 = 0.5$ and considering three different values for ε_2: 0.1, 0.25, 0.5.

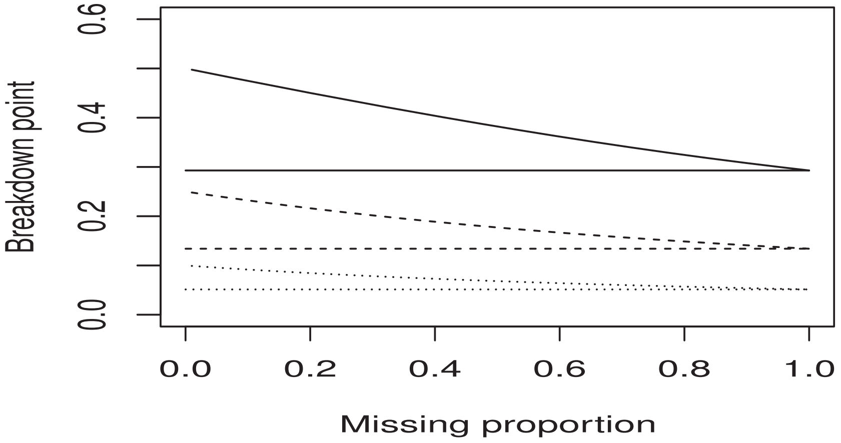

Figure 1. Breakdown points lower bounds for both procedures as a function of the missing fraction δ_n when $\varepsilon_1 = 0.5$ and three different values of ε_2. Solid lines correspond to $\varepsilon_2 = 0.5$, dashed lines for $\varepsilon_2 = 0.25$ and dotted lines are used for $\varepsilon_2 = 0.1$. In each case, constant lines represent $\varepsilon_3^{\text{SY}}$ while the others indicate how $\varepsilon_3^{\text{NEW}}$ varies as a function of δ_n.

5. Simulation study

5.1. Monte Carlo

In this section we present the results of a simulation study where we compare the performance of the estimation procedure introduced in this work with two previous proposals for estimation in presence of missing data: (i) the estimators presented in SY and (ii) those considered by Bianco et al. (2010) (BBMG, from now on). The simulation contemplates the following cases, also considered in SY for linear regression models.

Model 1 The response y is generated as $y = 5x_1 + x_2 + x_3 + 4v + 9$, where (x_1, x_2, x_3, v) is distributed according to a multivariate standard normal law.

Model 2 The response y is generated as $y = 5x_1 + x_2 + x_3 + 4v + 4$, where x_1, x_2 and x_3 are distributed according to a chi-squared distribution with one degree of freedom, and v has a standard normal distribution and the four variables are independent.

In both scenarios the errors have symmetric distribution. In the first case, also the response is symmetric, while in the second model the distribution of y is asymmetric.

The missing mechanism considered in both cases is the same. The conditional distribution on $\mathbf{x} = (x_1, x_2, x_3)$ of the variable a, which indicates that the response y is observed, satisfies

$$\ln\left(\frac{\mathrm{P}(a = 1|\mathbf{x})}{1 - \mathrm{P}(a = 1|\mathbf{x})}\right) = 0.15(x_1 + x_2 + x_3) \tag{16}$$

The mean value of the observed indicator is $\mathrm{P}(a = 1) = 0.5$ for Model 1 and $\mathrm{P}(a = 1) = 0.606$ for Model 2.

As in SY, we compute the estimators in two cases: with and without outliers. To study the behavior of the estimators under outlier contamination, we consider several scenarios, combining two amounts of contaminations ($\varepsilon = 0.1$ and $\varepsilon = 0.2$) and different values for the outliers. The contamination is performed according to the following scheme: $[\varepsilon m]$ of the m observations $(\mathbf{x}, y)$ (corresponding to $a = 1$) are replaced by the same value $(\mathbf{x}^*, y^*)$, and in $[\varepsilon(n - m)]$ of the remaining $n - m$ observations with $a = 0$, $\mathbf{x}$ is replaced by $\mathbf{x}^*$. Note that we use $[\cdot]$ to denote the integer part of a real number. We consider two values for $\mathbf{x}^*$, one corresponding to low leverage outliers, with $\mathbf{x}^* = (2, 0, 0)$, and for large leverage outliers we choose $\mathbf{x}^* = (10, 0, 0)$; y^* takes values in the intervals $[-40, 60]$ and $[-100, 200]$, when $\mathbf{x}^* = (2, 0, 0)$ and $\mathbf{x}^* = (10, 0, 0)$, respectively.

For each scenario we performed $N_{rep} = 1000$ replications using samples of size $n = 100$. We consider estimators for the mean (MEAN), and the parameters defined by four location and one dispersion continuous functionals. The location functionals are the median (MED), the 0.1-trimmed mean (TR10), an M-functional based on a Tukey loss function (TU) and an M-functional based on a Huber loss function (HU). M-location functionals are defined by

$$T(F) = \arg\min_{\mu\in\mathbb{R}} \mathrm{E}_F \left[\rho\left(\frac{y-\mu}{S(F)} \right) \right]$$

where $\rho(u)$ is a non decreasing function of $|u|$, and $S(F)$ is a dispersion functional. A brief review on M-location functionals can be found in Sec. 7.2 in SY, and a detailed presentation is included in Maronna, Martin and Yohai (2006) or in Huber and Ronchetti (2009). For TU, we use a function ρ in the Tukey bisquare family

$$\rho_{T,k}(u) = 1 - \left(1 - \left(\frac{u}{k}\right)^2\right)^3 I(|u| \le k) \tag{17}$$

with $k_1 = 3.44$. Instead, for HU ρ is taken in the Huber family

$$\rho_{H,k}(u) = u^2 I(|u| \le k) + (2k|u| - k^2) I(|u| > k)$$

with $k_2 = 1.37$. In both cases, the dispersion considered is defined through the dispersion functional $S(F)$, as follows. For each μ let $S^*(F, \mu)$ be the value solving

$$\mathrm{E}_F \left[\rho_0 \left(\frac{y-\mu}{S^*(F,\mu)} \right) \right] = \delta$$

Then $S(F)$ is given by

$$S(F) = \min_{\mu} S^*(F, \mu) \tag{18}$$

The function ρ_0 is taken in the Tukey bisquare family with $k = 1.547$ and $\delta = 0.5$. Finally, the dispersion functional to be estimated is given by $S(F_0)$, defined in (18) and we use (SS) to refer to it.

Note that the distribution F_0 of y derived under Model 1 is symmetric and, therefore, all the location functionals are equal to its center of symmetry, which is 9. Instead, under Model 2 all the location functionals take different values at F_0. More specifically, the values of the MED, TR10, TU, and HU at F_0 are given by 9.53, 10.07, 9.35, and 10.06, respectively. The dispersion $S(F_0)$ is 6.56 in Model 1 and 6.44 in Model 2.

The regression model was fitted with two different estimators using in both cases the subsample where the response variable is observed. When interested in the estimation of the mean, we take as $\widehat{\boldsymbol{\beta}}_n$ the least squares (LS) regression estimator. Instead, for the estimation of the parameters defined by continuous functionals, (MED, TR10, BI, HU, and SS), an MM-regression estimator is used. MM-estimators were introduced by Yohai (1987) to combine the highest possible breakdown point with an arbitrarily high efficiency for Gaussian errors, among equivariant estimators. As in the location case, these estimators minimize a ρ-function of the residuals, standardized with an M-estimator of the scale of the residuals based on a preliminary estimator. As we explained in Sec. 3, we have excluded the intercept in model (3). However, to get consistent estimators of $\boldsymbol{\beta}_0$ without requiring symmetric errors, it is necessary to estimate an additional parameter, which can be naturally interpreted as an intercept or a center of the error distribution. The MM-estimator that we use in the Monte Carlo study is defined in the following two steps.

Estimation of the error scale. Let $m = \sum_{i=1}^{n} a_i$, then given $(\alpha, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}$, let $\widehat{\sigma}(\alpha, \boldsymbol{\beta})$ be the solution of the following equation

$$\frac{1}{m}\sum_{i=1}^{n} a_i \rho_{T,k_0}\left(\frac{y_i - \alpha - \boldsymbol{\beta}^{\mathrm{t}}\mathbf{x}_i}{\widehat{\sigma}(\alpha, \boldsymbol{\beta})}\right) = 0.5$$

where $\rho_{T,k}$ is the Tukey bisquare family defined in (17) and $k_0 = 1.547$. Then the error scale is estimated by

$$\widehat{\sigma} = \min_{(\alpha,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \widehat{\sigma}(\alpha, \boldsymbol{\beta})$$

Estimation of the regression parameters. Consider

$$(\widehat{\alpha}_n, \widehat{\boldsymbol{\beta}}_n) = \operatorname*{argmin}_{(\alpha,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \sum_{i,j} a_i \rho_{T,k_1}\left(\frac{y_i - \alpha - \boldsymbol{\beta}^{\mathrm{t}}\mathbf{x}_i}{\widehat{\sigma}}\right) \tag{19}$$

where $k_1 = 3.44$.

Once the regression coefficients are estimated, we compute the residuals $\widehat{u}_i = y_i - \widehat{\boldsymbol{\beta}}_n^{\mathrm{t}}\mathbf{x}_i$, for $i \in A$.

We consider three estimators for each robust location parameter. The first two are obtained evaluating the corresponding functional at $\widehat{F}_n$, defined in (5), and at $\widetilde{F}_n$ defined in (8). Each of these procedures will be indicated with the subscript NEW and SY, respectively. The third estimator corresponds to the proposal of Bianco et al. (2010) and is obtained evaluating now each functional at the empirical distribution of the predicted response $\widehat{\boldsymbol{\beta}}_n^{\mathrm{t}}\mathbf{x}_i + \widehat{\alpha}_n$. We will use BBMG to refer to these last estimators. For estimating $S(F_0)$ we use $S(\widehat{F}_n)$ and $S(\widetilde{F}_n)$ since BBMG only considers the location case.

Under each scenario, we compute the (empirical) bias and mean square error (MSE) over the $N_{rep} = 1000$ replications, for each estimator applied to samples of size $n = 100$. Namely, for each case we computed Bias and MSE according to the following formulas:

$$\textbf{Bias} = \frac{1}{N_{rep}}\sum_{j=1}^{N_{rep}} \widehat{\theta}_{*,n,j} - \theta_0 \ , \quad \text{MSE} = \frac{1}{N_{rep}}\sum_{j=1}^{N_{rep}} (\widehat{\theta}_{*,n,j} - \theta_0)^2 \tag{20}$$

where $\widehat{\theta}_{*,n,j}$, $j = 1, \ldots, N_{nrep}$ denotes the j replication of the procedure $*$ applied to a sample of size $n = 100$, and $*$ should be replaced by BBMG, SY and NEW for the results corresponding to BBMG, SY, and NEW, respectively. We start presenting the results corresponding to the non contamination case.

Under Model 1, the symmetry of both the distribution of the response y and that of the error term u guarantee the consistency of BBMG estimators. However, in Model 2 the distribution of the response y is not symmetric and therefore BBMG estimators are not guaranteed to be consistent. For instance, the asymptotic bias under Model 2 for the MED, TR10, TU, and HU are given by 0.79, 0.39, 0.93 and 0.61, respectively. This fact precludes its use for this model, and for this reason BBMG procedures are not included in the simulation study for Model 2. On the other hand, as we already mentioned, there is not an estimator for the dispersion parameter analogous to BBMG's proposal, and therefore no BBMG's procedure is considered in the simulation for estimating the dispersion $S(F_0)$, defined in (18). Note that $T(\widehat{F}_n)$ and

Table 1. Monte Carlo results without contaminations.

	Model 1		Model 2	
	MSE_0	Bias_0	MSE_0	Bias_0
MEAN	0.59	0.02	0.85	− 0.02
MED_{BBMG}	0.86	0.03	—	—
MED_{SY}	0.65	0.02	0.63	−0.09
MED_{NEW}	0.75	0.02	0.73	−0.05
$\text{TR10}_{\text{BBMG}}$	0.72	0.01	—	—
TR10_{SY}	0.61	0.02	0.67	−0.03
TR10_{NEW}	0.62	0.02	0.67	−0.02
TU_{BBMG}	0.76	0.01	—	—
TU_{SY}	0.64	0.02	0.65	−0.05
TU_{NEW}	0.66	0.02	0.68	−0.02
HU_{BBMG}	0.71	0.01	—	—
HU_{SY}	0.61	0.02	0.67	−0.04
HU_{NEW}	0.62	0.02	0.67	−0.03
Scale_{SY}	0.43	− 0.05	0.39	−0.11
$\text{Scale}_{\text{NEW}}$	0.48	− 0.04	0.51	−0.07

$T(\widetilde{F}_n)$ are consistent in both models for all the functionals studied in this work. In Table 1, we show the results corresponding to models 1 and 2 without contaminations.

In the first column of Table 1, we observe that, as expected, when both the distribution of y and the distribution of the regression error u are Gaussian, the classical procedure is the most efficient. Even so, the efficiency of the robust estimators is quite high. Instead, in the third column of Table 1 we note that for Model 2 the robust procedures perform better than the classical one. The reason is that when y is generated as in Model 2, its distribution has heavy tails. It is also shown that, for all the functionals, the estimators based on the proposal presented in SY are the ones with the smallest MSE, closely followed by the estimators proposed here.

Let's now turn to the results under contaminations. We plot the MSE as a function of the outlier size y^* for each (available) estimators under different scenarios. For all the functionals, the conclusions that can be draw for $\varepsilon = 0.1$ and $\varepsilon = 0.2$ are very similar and so we include here the figures corresponding to the $\varepsilon = 0.1$ case, while those for $\varepsilon = 0.2$ are shown in the Supplementary Material. In Figures 2 and 3, we present the MSE curves for the location functionals under Model 1, with low ($\mathbf{x}^* = (2, 0, 0)$) and high ($\mathbf{x}^* = (10, 0, 0)$) leverage, respectively.

In both cases, BBMG procedures present the smallest MSE values. These results are not surprising since BBMG'estimators are specifically designed for the case that the response y has a symmetric distribution, requirement that is satisfied by Model 1. Excepting BBMG, there is no estimator with uniform best behavior. The largest MSEs are attained when the outliers y^* are located on the right hand side of the horizontal axis. For these type of outliers, the new estimators compare favorably with respect to the SY'ones, except for the Tukey M-location functional. However, in this case, the least favorable situation corresponds to $\mathbf{x}^* = (2, 0, 0)$, where the MSE of the new estimator is smaller than that of the procedures presented in SY. Then, if we can not assume that y has a symmetric distribution and we evaluate the performance of each procedure through a minimax criterion based on the MSE, the new proposal is recommended; its advantage is neater for the median.

We observe that TR10 does not behave as good as the other robust estimators. This is also a result of the *outliers propagation*. Specifically, even if the regression estimator has a breakdown point $\varepsilon_2 = 0.5$ and the UABP of the 0.1-trimmed mean is $\varepsilon_1 = 0.1$, the lower bound

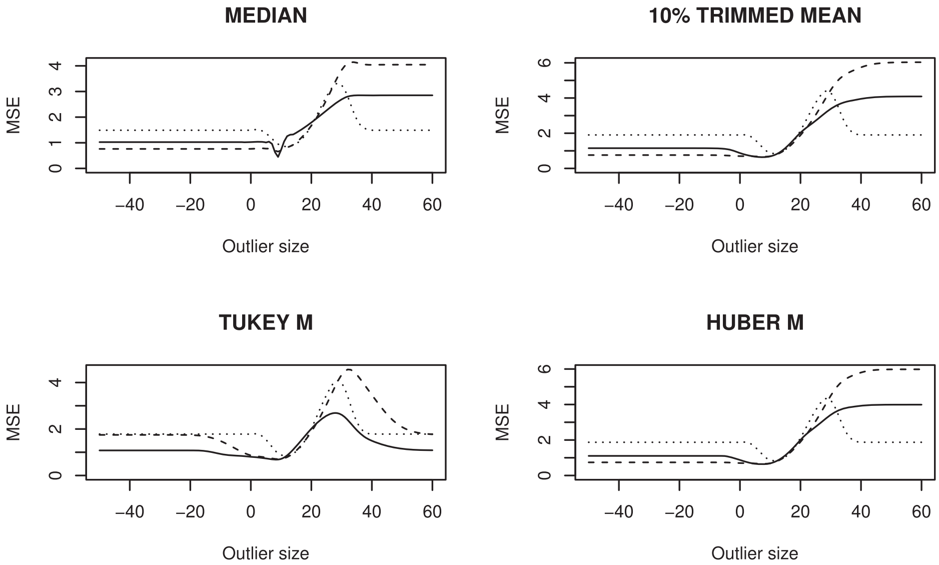

Figure 2. MSEs under outlier contamination for Model 1, with $\varepsilon = 0.1$ and $\mathbf{x}^* = (2, 0, 0)$. Dotted lines correspond to BBMG, dashed lines are used for SY, and solid lines represent the results corresponding to the new estimators.

$\varepsilon_3^{\text{NEW}}$ given in (15) for this estimator is 0.068 when $\eta_n = 0.5$ and 0.073 when $\eta_n = 0.6$. As a consequence this estimator cannot cope with 10% of outliers.

Figures 4 and 5 present MSEs for location functionals under Model 2, with low ($\mathbf{x}^* = (2, 0, 0)$) and high ($\mathbf{x}^* = (10, 0, 0)$) leverage, respectively. The comments presented for Model 1 remain valid for Model 2.

In Figure 6, we present the MSE curves for estimating the dispersion $S(F_0)$, with $\varepsilon = 0.1$. Model 1 and Model 2 are considered in the top and bottom panels, respectively. The

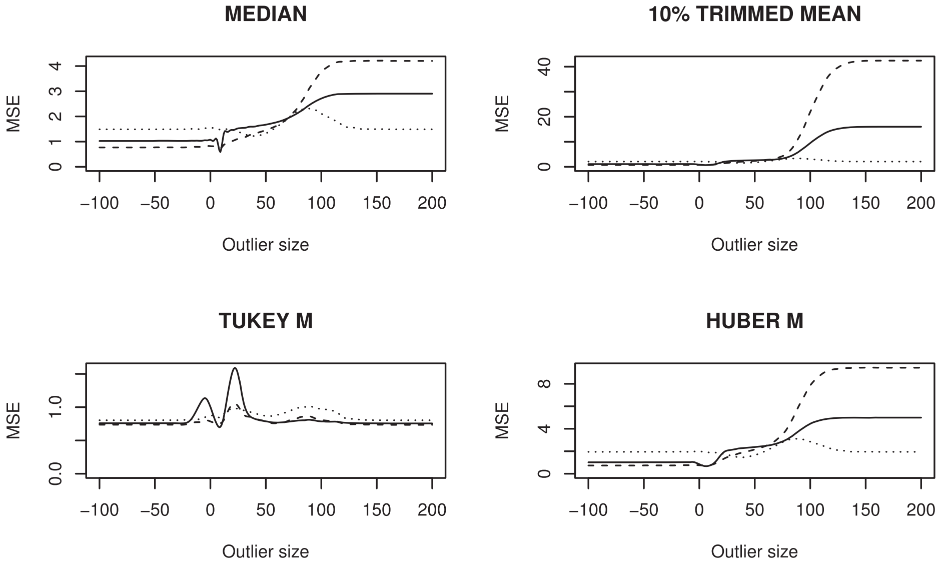

Figure 3. MSEs under outlier contamination for Model 1, with $\varepsilon = 0.1$ and $\mathbf{x}^* = (10, 0, 0)$. Dotted lines correspond to BBMG, dashed lines are used for SY, and solid lines represent the results corresponding to the new estimators.

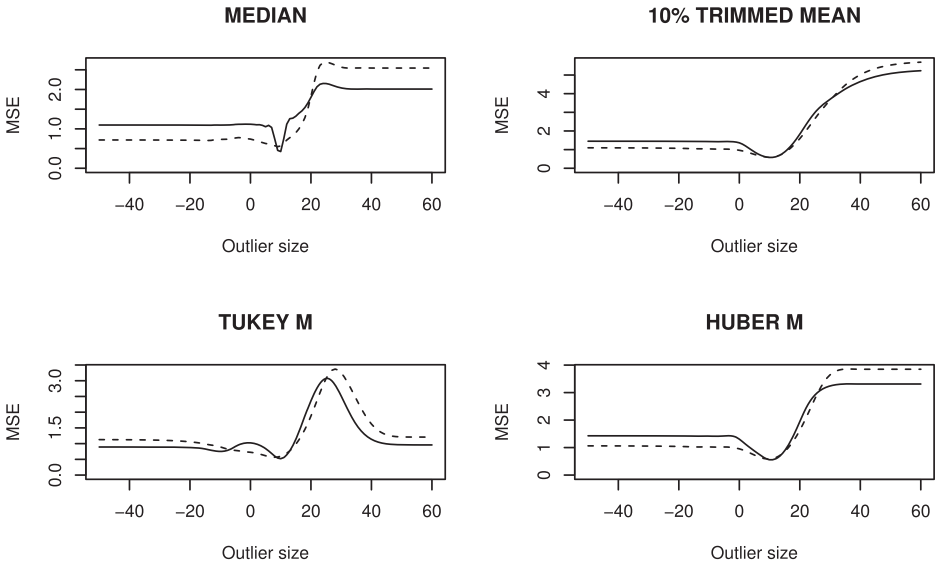

Figure 4. MSEs under outlier contamination for Model 2, with $\varepsilon = 0.1$ and $\mathbf{x}^* = (2, 0, 0)$. Dashed lines are used for SY and solid lines represent the results corresponding to the new estimators.

results corresponding to $\mathbf{x}^* = (2, 0, 0)$ are in the left panel while those corresponding to $\mathbf{x}^* = (10, 0, 0)$ are on the right panel. For Model 2 the maximum values of the MSE correspond to $\mathbf{x}^* = (10, 0, 0)$ and y^* on the right hand side of the horizontal axis. For this case, the new procedure is much better than the previous one. The MSE curves corresponding to the new estimators are, practically, uniformly below those of the SY estimator for $\varepsilon = 0.2$ (see Supplementary Material). This shows that the new procedure is clearly preferable for estimating the dispersion parameter $S(F_0)$, defined in (18).

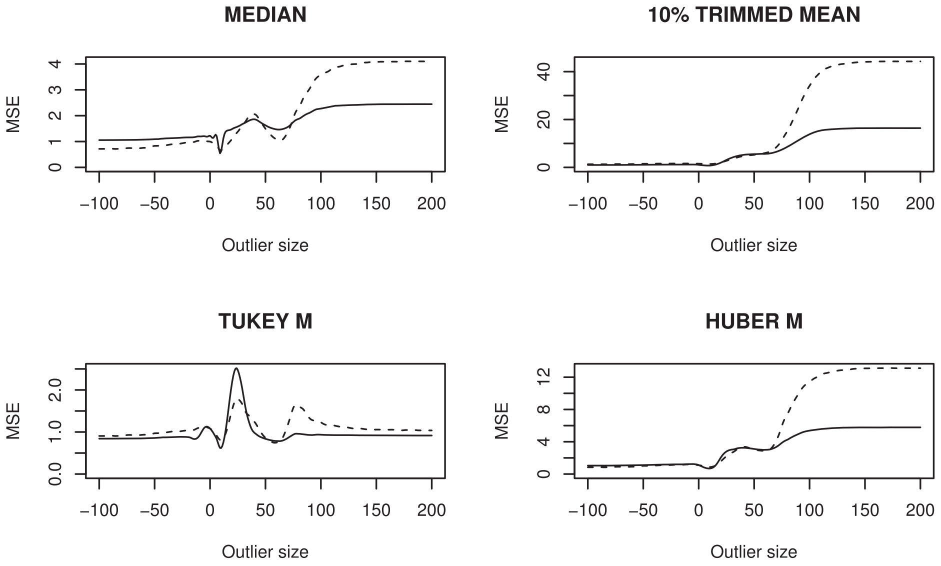

Figure 5. MSEs under outlier contamination for Model 2, with $\varepsilon = 0.1$ and $\mathbf{x}^* = (10, 0, 0)$. Dashed lines are used for SY and solid lines represent the results corresponding to the new estimators.

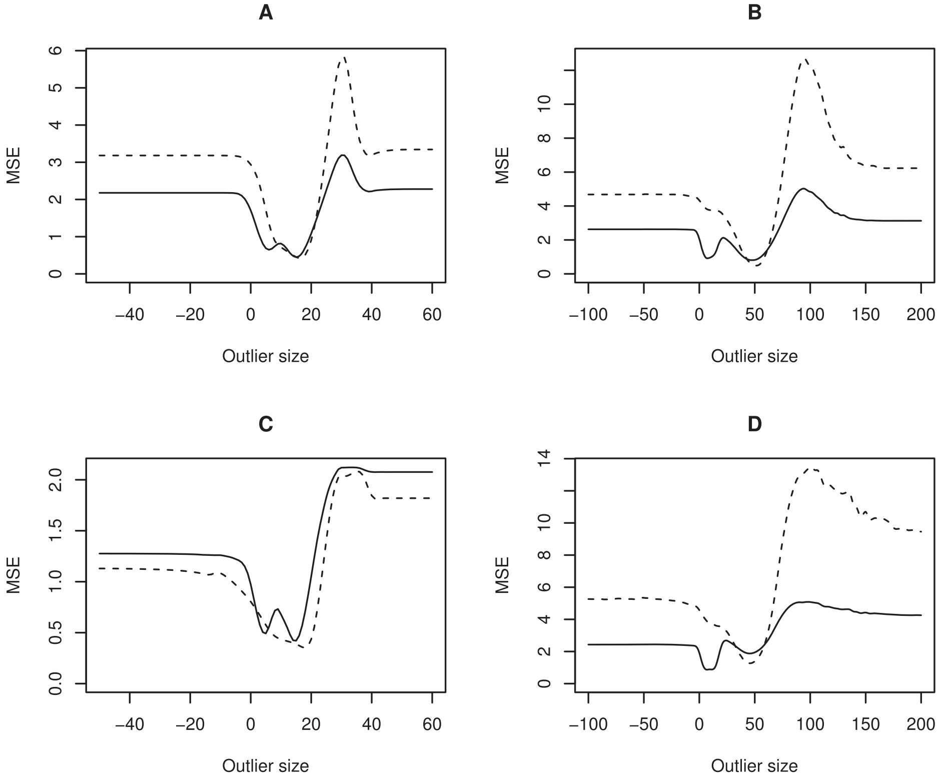

Figure 6. Dispersion: MSEs under outlier contamination for $\varepsilon = 0.1$ and (A) Model 1 and $\mathbf{x}^* = (2, 0, 0)$, (B) Model 1 and $\mathbf{x}^* = (10, 0, 0)$, (C) Model 2 and $\mathbf{x}^* = (2, 0, 0)$, and (D) Model 2 and $\mathbf{x}^* = (10, 0, 0)$. Dashed lines are used for SY and solid lines represent the results corresponding to the new estimators.

Tables 1 and 2 in the Supplementary Material show the maximum mean squared error ($\text{MSE}_{\max}$) under outliers contamination over all the considered values for y^*. They also contain the value of y^* where the maximum mean squared error is achieved (ymax). This is done for each amount ε of outliers, for each possible contamination $\mathbf{x}^*$, under Model 1 and Model 2, respectively.

A formal analysis of the computation times required for the different procedures is beyond the scope of this work. However, in Table 2 we present the mean time needed by the entire procedure to estimate the Tukey M-location functional. In this table we consider $p = 5$, 10, 20, 40 and different values of n; the missing proportion in each case is $\eta = 0.5$. This study has been done using our R code available in the Supplementary Material running on an Intel Core i7 (3.60 GHz) processor on Windows 7 operative system. These times can be reduced by computing more efficiently the location estimators, using weights instead of the entire sample of pseudo observations, given that many of them are repeated.

Table 2. Times required for computing the estimators.

	$p = 5$		$p = 10$		$p = 20$		$p = 40$	
Computer times	$n = 50$	$n = 100$	$n = 100$	$n = 200$	$n = 200$	$n = 400$	$n = 400$	$n = 800$
t_{SY}	0.01	0.04	0.07	0.18	0.14	0.56	0.70	3.08
t_{NEW}	0.01	0.04	0.08	0.19	0.13	0.57	0.67	3.09

Table 3. Real data analysis.

	MED	TR10	BI	HU	SCALE
$y_i, 1 \leq i \leq 297$	4232.31	4910.59	4234.29	5007.08	4714.53
$\text{MSE}_{\text{OTHER}}$	116898.20	647948.73	116537.47	817429.08	
MSE_{SY}	55457.24	40323.07	67422.55	45590.85	46910.68
MSE_{NEW}	48828.94	39547.17	76726.39	50420.76	42099.16

MSE based on $Nrep = 1000$ data sets constructed from the original one by artificially removing responses, as indicated in model (21).

5.2. Real data example

To analyze the performance of our proposal in a real data example, we consider a data set where there are not missing responses. Then, we generate $Nrep = 1000$ data sets with artificially missing responses by removing some of them. Finally, we evaluate the performance of the different estimators.

The data set is composed of $n = 297$ individuals and the response y of interest is the annual salary corresponding to 1978. This data set is part of a study published by LaLonde (1986), who was interested in evaluating the impact of a training program on the salaries. We restrict our attention to the trained set. A complete description of this data set can be found in LaLonde (1986), while data are available at http://users.nber.org/~rdehejia/data/nsw_treated.txt.

For each individual a vector $\mathbf{x}$ of seven covariates is observed. We consider the same continuous functionals used in the simulation study: the median (MED), the 10% trimmed mean (TR10), the Tukey M-location (TU), the Huber M-location (HU) and the dispersion defined in (18) (SS). The values of each functional at the empirical distribution of the responses are shown in the first line of Table 3. To emulate a missing data setting, we generate $Nrep = 1000$ data sets with artificially missing responses. To do so, in each replication, we keep the covariates $\mathbf{x}_i$ $(i = 1, \ldots, 297)$ and remove some of the responses y_i. For each of these data sets, the observed indicator a is generated according to the following mechanism, proposed in SY:

$$\ln\left(\frac{\mathrm{P}(a = 1|\mathbf{x})}{1 - \mathrm{P}(a = 1|\mathbf{x})}\right) = 0.001x_7 \tag{21}$$

where x_7 represents the earnings corresponding to 1975. At each replication, the responses y_i with $a_i = 0$ are considered missing. For each data set we compute all the estimators considered in the Monte Carlo study: those proposed in SY, the procedures presented in the present work and also the proposals of Bianco et al. (only for location functionals). Then, we compute mean square errors, as presented in Equation (20), but replacing θ_0 by the corresponding functional evaluated at $\widehat{F}_y$ (values presented in line 1 of Table 3). Finally, we perform a paired t test to evaluate the significance of the observed MSE differences between the proposal presented in SY and the new one (lines 3 and 4). All the MSE are significantly different ($p < 0.001$), except for the TR10 functional. The bad behavior of Bianco et al. (2010) can be attributed to the fact that the response distribution is highly asymmetric (see density plot in Figure 6 of the Supplementary Material). We should highlight that in the case of the median, which is one of the most popular location parameters, the new estimator behaves significantly better than that proposed in SY.

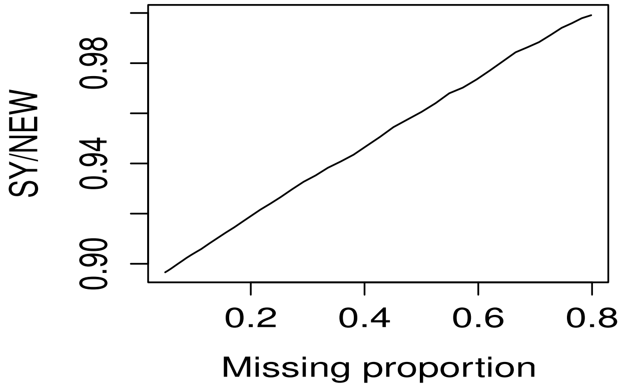

Figure 7. Efficiency: the ratio $\tau^2_{\text{SY}}(\xi)/\tau^2_{\text{NEW}}(\xi)$ is plotted as a function of the missing data ratio $\delta(\xi)$.

6. Numerical comparison of asymptotic variances.

The asymptotic variance τ^2 of the estimator $\widehat{\theta}_n = T(\widehat{F}_n)$, with $\widehat{F}_n$ defined in (5), is exhibited in equation (11). Even if η (the proportion of observed responses) is explicitly present in formula (11), we want to emphasize that τ^2 depends on the joint distribution of the vector $(\mathbf{x}^t, a, y)$. This is also the case for the estimators presented in SY. However, for a better understanding of the proposals, we compute asymptotic variances under scenarios with different ratio of missing data $\delta = 1 - \eta$. Namely, we generate $(\mathbf{x}^t, y)$ as in Model 1, but now the probability of $a = 1$ given $\mathbf{x} = (x_1, x_2, x_3)$ is given by

$$\ln\left(\frac{\text{P}(a=1|\mathbf{x})}{1-\text{P}(a=1|\mathbf{x})}\right) = 0.15(x_1 + x_2 + x_3) + \xi \tag{22}$$

In this way, for each ξ, the missing ratio is $\delta(\xi) = 1 - \eta(\xi)$, where $\eta(\xi) = \text{E}[\text{expit}\{0.15(x_1 + x_2 + x_3) + \xi\}]$, while $\text{expit}(u) = e^u/(1+e^u)$. We vary ξ in such a way that the missing ratio $\delta(\xi)$ takes values between 0.05 and 0.8. Let T_{TU} be the Tukey M-location functional, as defined in the Monte Carlo study. Consider the estimators $T_{\text{TU}}(\widetilde{F}_n)$ and $T_{\text{TU}}(\widehat{F}_n)$ and let $\tau^2_{\text{SY}}(\xi)$ and $\tau^2_{\text{NEW}}(\xi)$ denote the asymptotic variance of each estimator when $(\mathbf{x}^{\text{t}}, a, y)$ satisfies both Model 1 and (22). An explicit formula for τ^2_{NEW} is presented in equation (11) while in SY we present a formula for τ^2_{SY}. The efficiency of $T_{\text{TU}}(\widetilde{F}_n)$ with respect to $T_{\text{TU}}(\widehat{F}_n)$ is defined by

$$\text{EFF}(\xi) = \frac{\tau^2_{\text{SY}}(\xi)}{\tau^2_{\text{NEW}}(\xi)} \tag{23}$$

In Figure 7, we plot $(\delta(\xi), \text{EFF}(\xi))$.

We see that, when the missing ratio δ is small, the variance of the new proposal is slightly larger than that of the estimator presented in SY. The reason of the larger efficiency in SY can be attributed to the fact that in this procedure all the responses, even the observed ones, are predicted using a correct specified model for the regression. As the missing ratio increases, the difference in efficiency between the two proposals becomes negligible.

7. Conclusions

We have presented a new procedure to estimate the distribution of a variable y when there are missing data, which gives rise to new estimators for parameters defined through continuous functionals. A simulation study and a real data analysis show that the new procedure is highly robust and efficient, except for the symmetric case, where BBMG's estimators for location parameters are preferable.

In particular, we highlight the good performance of the new estimator for the median and for the dispersion.

Finally, we should mention the possibility of extending the result of this work to the generalized linear model (GLM), where, instead of considering the regression model presented in (3), we can assume that the distribution of y_i given $\mathbf{x}_i$ follows a GLM. That is, $y_i|\mathbf{x}_i \sim H_{g(\mathbf{x}_i^{t}\beta_0),\sigma_0}$, $1 \leq i \leq n$, where $H_{\theta,\sigma}$ is a parametric family of univariate distributions. In this case, we can estimate F_0 by $\widehat{F}_n(y) := n^{-1}\sum_{i=1}^{n} H_{g(\mathbf{x}_i^{t}\widehat{\beta}_n),\widehat{\sigma}}(y)$, where $\widehat{\beta}_n$ and $\widehat{\sigma}$ are consistent estimators for β_0 and σ_0, respectively, computed with $(\mathbf{x}_i, y_i)$ such that $a_i = 1$. The analysis of the properties of these procedures are part of a work in progress.

Acknowledgments

The authors wish to thank the Associate Editor and two anonymous referees for valuable comments which led to an improved version of the original paper. This research was partially supported by Grants pict 2014-0351 from anpcyt, 20020120200244ba and 20020130100279BA from the Universidad de Buenos Aires at Buenos Aires, Argentina.

Supplementary material

A Supplementary Material is available with the proofs of the theoretical results, the figures and tables corresponding to Sec. 5 and a code with the functions used in the simulation study.

References

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61 (4):962–73.

Bianco, A., G. Boente, W. González-Manteiga, and A. Pérez-González. 2010. Estimation of the marginal location under a partially linear model with missing responses. *Computational Statistics & Data Analysis* 54:546–64.

Donoho, D. L., and P. J. Huber. 1983. The notion of breakdown point. In *A festschrift for E. L. Lehmann*, eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges, 157–84. Belmont, CA: Wadsworth.

Fasano, M. V., R. A. Maronna, M. Sued, and V. J. Yohai. 2012. Continuity and differentiability of regression M functionals. *Bernoulli* 4:1284–309.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69:383–93.

Hájek, J. 1968. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics* 325–46.

Huber, P. J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35:73–101.

Huber, P. J., and E. M. Ronchetti. 2009. *Robust statistics*. 2nd ed. New York: Wiley.

LaLonde, R. J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76:604–20.

Maronna, R. A., R. D. Martin, and V. J. Yohai. 2006. *Robust statistics: Theory and methods*. Chichister: Wiley.

Rotnitzky, A., Q. Lei, M. Sued, and J. M. Robins. 2012. Improved double-robust estimation in missing data and causal inference models. *Biometrika* 99 (2):439–56.

Rubin, D. 1976. Inference and missing data. *Biometrika* 63:581–92.

Rousseeuw, P. J., and V. J. Yohai. 1984. Robust regression by means of S-estimators. In *Robust and nonlinear time series*, Lectures Notes in Statistics, eds. J. Franke, W. Hardle, and R. D. Martin, vol. 26, 256–272. New York: Springer.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5):688.

Sued, M., and V. J. Yohai. 2013. Robust location estimation with missing data. *Canadian Journal of Statistics* 41:111–32. Supplemental Material available online.

Yohai, V. J. 1987. High breakdown–point and high efficiency estimates for regression. *The Annals of Statistics* 15:642–56.

Westreich, D., J. K. Edwards, S. R. Cole, R. W. Platt, S. L. Mumford, and E. F. Schisterman. 2015. Imputation approaches for potential outcomes in causal inference. *International Journal of Epidemiology* 44 (5):1731–7.

Zhang, Z., Z. Chen, J. F. Troendle, and J. Zhang. 2012. Causal inference on quantiles with an obstetric application. *Biometrics* 68:697–706.