A new estimator for a finite population cdf in presence of auxiliary information

Leo Pasquazzi and Lucio De Capitani

Abstract In this work we propose a new estimator for the finite population cdf of a study variable that combines the two approaches to exploit knowledge about an auxiliary variable used in the Chambers and Dunstan ([2]) and the Kuo ([5]) estimators. As both the latter estimators, the new estimator is based on a superpopulation model where the population values of the study variable are generated independently from a model-cdf that is allowed to depend smoothly on an auxiliary variable. Like the Chambers and Dunstan estimator, the new estimator is based on estimates for the model-cdf of the study variable that are obtained by estimating the model-mean regression function and the model-cdf of the error terms separately. In the new estimator however both estimation steps are performed by non parametric regression in order to account for superpopulation models with smooth mean regression function and error term distribution that depends smoothly on the auxiliary variable. The non parametric regression for estimating model-cdf of the error terms resembles the one used in the Kuo estimator to estimate the model-cdf of the study variable directly, without considering the model-mean regression function. We will present a simulation study which shows that the new estimator outperforms several well known estimators from literature when the error terms are independently but not identically distributed.

Leo Pasquazzi
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milan, e-mail: leo.pasquazzi@unimib.it

Lucio De Capitani
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milan, e-mail: lucio.decapitani1@unimib.it

1 Introduction

Virtually all research articles that deal with the problem of estimating a distribution function in the context of finite population sampling quote the estimator proposed by Chambers and Dunstan ([2]). In fact, the article written by these two authors appears to provide the historically first attempt to motivate a procedure for estimating the finite population cdf of a study variable that takes advantage of knowledge about an auxiliary variable. The original definition of the Chambers and Dunstan estimator (henceforth CD-estimator) is based on a superpopulation model where the population values of the study variable are given by

$$y_i := x_i\beta + v(x_i)\varepsilon_i, \qquad i = 1, 2, \ldots, N, \tag{1}$$

where (i) $x_1, x_2, \ldots x_N$ are known values taken on by some auxiliary variable, (ii) β is an unknown parameter, (iii) $v(\cdot)$ is a known function, and (iv) the error terms ε_i are i.i.d. random variables with $E\{\varepsilon_i\} = 0$ and $Var\{\varepsilon_i\} = \sigma^2$ for $1 \leq i \leq N$. The purpose of the CD-estimator is to estimate the finite population cdf of the study variable having observed the values taken on by the study variable corresponding to a sample $s \subset \{1, 2, \ldots, N\}$ of index values i, that has been chosen independently from the random mechanism that generated the population values of the study variable. Denoting as usual by $I(\cdot)$ the indicator function that takes on the value 1 if the statement in its argument is true, and is equal to 0 otherwise, the finite population cdf of the study variable is defined by

$$F_N(t) := \frac{1}{N}\sum_{i=1}^{N} I(y_i \leq t), \qquad t \in \mathbb{R}.$$

Once the values y_i are observed for $i \in s$, the sample cdf

$$F_n(t) := \frac{1}{n}\sum_{j \in s} I(y_j \leq t)$$

will be known, and thus the CD-estimator, as every model-based estimator for $F_N(t)$, may be written as

$$\widehat{F}(t) := \frac{1}{N}\left\{nF_n(t) + (N-n)\widehat{F}_{N-n}(t)\right\}$$

where $\widehat{F}_{N-n}(t)$ is a predictor for the non-sample cdf

$$F_{N-n}(t) := \frac{1}{N-n}\sum_{i \notin s} I(y_i \leq t).$$

In the CD-estimator $\widehat{F}_{N-n}(t)$ is given by

$$\widehat{F}_{CD,N-n}(t) := \frac{1}{N-n} \sum_{i \notin s} \widehat{G}_{CD} \left(\frac{t - x_i \widehat{\beta}}{v(x_i)} \right)$$

where

$$\widehat{G}_{CD}(\varepsilon) := \frac{1}{n} \sum_{j \in s} I \left(\frac{y_j - x_j \widehat{\beta}}{v(x_j)} \leq \varepsilon \right) \tag{2}$$

is an estimator for the model-cdf $G(\varepsilon)$ of the error terms, with

$$\widehat{\beta} := \frac{\sum_{j \in s} v^{-2}(x_j) x_j y_j}{\sum_{j \in s} v^{-2}(x_j) x_j^2}$$

the weighted least squares estimator for the unknown value of the parameter β.

The main drawbacks of the original version of the CD-estimator stem from the fact that its definition assumes a linear model-mean regression function and that it requires the user to specify the function $v^2(\cdot)$ that describes the variance of the error terms. In fact, as Chambers and Dunstan ([2]) themselves point out, both these problems are quite critical, since any deviation from model (1) introduces an asymptotic model-bias. A modified version of the CD-estimator, that accommodates also nonlinear model-mean regression functions, has been analyzed by Dorfman and Hall ([3]), who considered a superpopulation model where

$$y_i := \mu(x_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, N, \tag{3}$$

with (i) $\mu(\cdot)$ a smooth function, and (ii) ε_i i.i.d. error terms. Dorfman and Hall (1993) estimate $\mu(\cdot)$ by

$$\widehat{\mu}(x) := \sum_{j \in s} w_j(x) y_j,$$

using Gaussian kernel weights in place of $w_j(\cdot)$ and use

$$\widehat{F}_{DH,N-n}(t) := \frac{1}{N-n} \sum_{i \notin s} \widehat{G}_{DH} \left(t - \widehat{m}(x_i) \right)$$

with

$$\widehat{G}_{DH}(\varepsilon) := \frac{1}{n} \sum_{j \in s} I(y_j - \widehat{m}(x_i) \leq \varepsilon)$$

as predictor for the non-sample cdf $F_{N-n}(t)$. Below the resulting estimator for $F_N(t)$ will be called DH-estimator.

Kuo ([5]) proposed another approach to estimate $F_N(t)$ based on a superpopulation model where it is merely assumed that

$$P\{y_i \leq t\} := p(x_i), \qquad i = 1, 2, \ldots, N, \tag{4}$$

for some smooth function $p(\cdot)$. In Kuo-type estimators, the implied predictor for the non-sample cdf $F_{N-n}(t)$ is given by

$$\widehat{F}_{K,N-n}(t) := \sum_{i \notin s} \widehat{p}_K(x_i),$$

where

$$\widehat{p}_K(x) := \sum_{j \in s} w_j(x) I(y_j \leq t) \tag{5}$$

is a non-parametric regression estimator for $p(x)$ based on the regression weights $w_j(x)$. Kuo ([5]) suggested three types of weights $w_j(\cdot)$: (i) uniform kernel weights, (ii) gaussian kernel weights and (iii) nearest k-neighbor weights.

It is worth noting that the CD-estimator and Kuo-type estimators employ different kind of estimators for $P\{y_i \leq t\}$ for $i \notin s$. While the CD-estimator uses

$$\widehat{p}_{CD}(x_i) := \widehat{G}_{CD}\left(\frac{t - x_i\widehat{\beta}}{v(x_i)}\right),$$

which may be viewed as a semiparametric estimator, Kuo-type estimators use the completely nonparametric estimator $\widehat{p}_K(x_i)$. One would obviously expect that $\widehat{p}_{CD}(x_i)$ is more efficient than $\widehat{p}_K(x_i)$ if model (1) is true, and that the opposite holds otherwise. In fact, it may be shown that $\widehat{p}_{CD}(x_i)$ is asymptotically model-biased for $p(x_i)$ unless model (1) is true, while $\widehat{p}_K(x_i)$ is asymptotically model-unbiased under the much more general superpopulation model (4). To overcome this shortcoming of $\widehat{p}_{CD}(x_i)$ one might consider

$$\widehat{p}_{CDW}(x_i) = \widehat{p}_{CD}(x_i) + \sum_{j \in s} w_j(x_i)\left[\widehat{p}_K(x_j) - \widehat{p}_{CD}(x_j)\right]$$

as estimator for $P\{y_i \leq t\}$, and thus

$$\widehat{F}_{CDW,N-n}(t) := \sum_{i \notin s} \widehat{p}_{CDW}(x_i)$$

as predictor for $F_{N-n}(t)$ in model-based estimators for $F_N(t)$. The resulting estimator for $F_N(t)$ turns out to be the one Chambers, Dorfman and Wherly ([1]) end up with when applying their bias correction procedure to the problem of estimating $F_N(t)$. We therefore call it CDW-estimator in what follows. Dorfman and Hall ([3]) analyzed the CDW estimator, with Gaussian kernel weights in place of $w_j(\cdot)$. They showed that its model-bias achieves the parametric $O(n^{-1})$ rate (like that of the CD-estimator) when model (1) is true, and that the convergence rate of the model-bias is of order $O(\lambda^2) + o((n\lambda)^{-1})$ (like that of the Kuo estimator with Gaussian kernel weights), where λ denotes the bandwidth in the gaussian kernel weights, under the more general model (4) when model (1) is false. Dorfman and Hall ([3]) showed moreover that the convergence rate of the model-variance of the CDW-etimator is $O(n^{-1})$ under model (4) (like that of the CD, DH and Kuo estimator), irrespective of whether model (1) is true or false.

In this work we shall explore still another approach to estimate $P\{y_i \leq t\}$ for use in predictors of $F_{N-n}(t)$. The idea is to use nonparametric regression weights to estimate first the model-mean regression function $\mu(\cdot)$, and then, using the regression residuals, the model-cdfs $G(\cdot,\cdot)$ of the error terms. The model-cdfs of the error terms have a second argument because they may depend on the auxiliary variable. The estimator for $P\{y_i \leq t\}$ we propose is thus defined by

$$\widehat{p}_{CDK}(x_i) := \sum_{j\in s} w_j(x_i) I\left(y_j - \widehat{\mu}(x_i) \leq t\right).$$

The regression weights $w_j(\cdot)$ for estimating $\mu(\cdot)$ and $G(\cdot,\cdot)$ may actually be different. Since $\widehat{p}_{CDK}(x_i)$ is based on the ideas underlying the definitions of both $\widehat{p}_{CD}(x_i)$ and $\widehat{p}_K(x_i)$, we used the acronym CDK (Chambers-Dunstan-Kuo) in the notation and we call the corresponding model-based estimator for $F_N(t)$ CDK-estimator in what follows.

The above approaches to estimate $P\{y_i \leq t\}$ may of course be used to define model-assisted estimators for $F_N(t)$ as well (see [4]). Model assisted estimators for $F_N(t)$ are estimators of the form

$$\widetilde{F}(t) := \frac{1}{N}\left\{\sum_{i=1}^{N} \widetilde{p}_i + \sum_{i\in s} \pi_i^{-1}\left[I(y_i \leq t) - \widetilde{p}_i\right]\right\}$$

where $\widetilde{p}_1, \widetilde{p}_2, \ldots, \widetilde{p}_N$ are fitted values for $I(y_1 \leq t), I(y_2 \leq t), \ldots, I(y_N \leq t)$ based on some superpopulation model, and $\pi_1, \pi_2, \ldots, \pi_N$ are the first order inclusion probabilities corresponding to the sample design by which the sample s has been drawn. The fitted values $\widetilde{p}_i$ may be obtained by adapting the definitions of $\widehat{p}_{CD}(x_i)$, $\widehat{p}_{DH}(x_i)$, $\widehat{p}_{CDW}(x_i)$ and $\widehat{p}_{CDK}(x_i)$ to account for the sample design.

The advantage of model-assisted estimators lies in the fact that they are asymptotically design-unbiased even if the assumed superpopulation model is false, and thereby they provide protection against model misspecification. However, this advantage comes at the cost of larger asymptotic design-variance in comparison with model-based estimators. Model-based estimators will thus be more efficient than model-assisted ones if their design-bias is not to large. This will be the case if the data support the superpopulation model and if the first order inclusion probabilities do not vary a lot.

In the following section we will compare design-bias and design-variance of the above estimators in a simulation study. Since we consider simple random without replacement sampling, we use $\widehat{p}_{CD}(x_i)$, $\widehat{p}_{DH}(x_i)$, $\widehat{p}_{CDW}(x_i)$ and $\widehat{p}_{CDK}(x_i)$ as they are to compute the fitted values $\widetilde{p}_i$ in the model assisted estimators. As for the nonparametric regression weights $w_j(\cdot)$, we always use local linear regression weights, defined by

$$w_j(x) := \frac{1}{n\lambda} K\left(\frac{x - x_j}{\lambda}\right) \frac{M_{n2}(x) - (x - x_j) M_{n1}(x)}{M_{n2}(x) M_{n0}(x) - M_{n1}(x)^2}, \qquad x \in \mathbb{R},$$

where $K(\cdot)$ is the Epanechnikov kernel function,

$$M_{nk}(x) := \frac{1}{n\lambda}\sum_{j\in s} K\left(\frac{x-x_j}{\lambda}\right)(x-x_j)^k, \qquad k=0,1,2,$$

and $\lambda > 0$ is the bandwidth parameter.

2 Simulation Results

In this section we compare the performance of the above estimators for $F_N(t)$ in a simulation study. We consider finite populations of size $N = 1000$ generated from the following superpopulation models:

$$y_i := 0.5x_i + \varepsilon_i$$
$$y_i := \sqrt{x_i} + \varepsilon_i$$

where ε_i are independent error terms generated either from the Student t distribution with $\nu = 5$ dgf, or from shifted noncentral Student t distributions with $\nu = 5$ dgf and with noncentrality parameter given by $\zeta = 15x_i$. The shifts applied to the error term distributions in the second case are aimed to make sure that their expectation equals zero. The x-values taken on by the auxiliary variable will be generated independently from the uniform distribution on $(0,1)$.

The populations we consider are thus 4: the first one with linear regression function and i.i.d. error terms (LRIDE-population), the second one with linear regression function and independent non identically distributed error terms (LRNIDE-population), the third one with nonlinear regression function and i.i.d. error terms (NLRIDE-population) and the last one with nonlinear regression function and independent but non identically distributed error terms (NLRNIDE-population).

To compare the performance of the estimators, we select $B = 1000$ samples from each of the four populations by simple random without replacement sampling and evaluate the estimation error of each cdf-estimator at $t_k = F_N^{-1}(k/20)$ for $k = 1,2,\dots,19$. In the estimators that involve nonparametric regression weights $w_j(\cdot)$ we use local linear regression weights with Epanechnikov kernel function. As for the bandwidth, we test three values: $\lambda = 0.1$, $\lambda = 0.2$ and $\lambda = 0.3$. Since the simulation results are rather insensitive to the bandwidth, we report only the results referring to $\lambda = 0.1$ in Figures 2 to 8. The latter show the population values of the study variable plotted against the auxiliary variable and the simulated bias, standard deviation and rmse plotted against $p = k/20$. In the legend, ”XXma” identifies the model-assisted version of the XX-estimator, while ”XXmb” identifies the corresponding model-based version. The graphs emphasize, as expected (see discussion in [4]), that the model-assisted estimators perform very similarly. They all are nearly unbiased, and feature very similar standard errors. The performance of the model-based estimators, on the other hand, is more diverse. The CD and DH-estimators are the most efficient ones in the populations they are designed for (the LRIDE for the CD-estimator, and the LRIDE and NLRIDE populations for the DH-estimator), but they are inefficient otherwise. The model-based CDK-estimator appears to be the most robust one. It is more efficient than the model-assisted estimators in all populations, and it does not suffer from model-misspecification bias like the CD and DH-estimators. The performance of the model-based CDW-estimator appears to be influenced mainly by that of the model-based KUO-estimator, which appears to be the less efficient one except in the populations where the model misspecification bias spoils the performance of the CD and DH-estimators.

References

1. Chambers, R.L., Dorfman, A.H., Wehrly, T. (1993). Bias robust estimation in finite populations using non-parametric calibration. *J. Amer. Statist. Assoc.*, 88(421), 268–277.
2. Chambers, R.L., Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597–604.
3. Dorfman, A.H., Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452–1475.
4. Johnson, A.A., Breidt, F.J., Opsomer, J.D. (2008). Estimating Distribution Functions from Survey Data Using Nonparametric Regression. *J. Stat. Theory Pract.*, 2(3), 419–431.
5. Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *Proceedings of the section on survey research methods* (pp. 280-285), Amer. Statist. Assoc., Alexandria, VA.

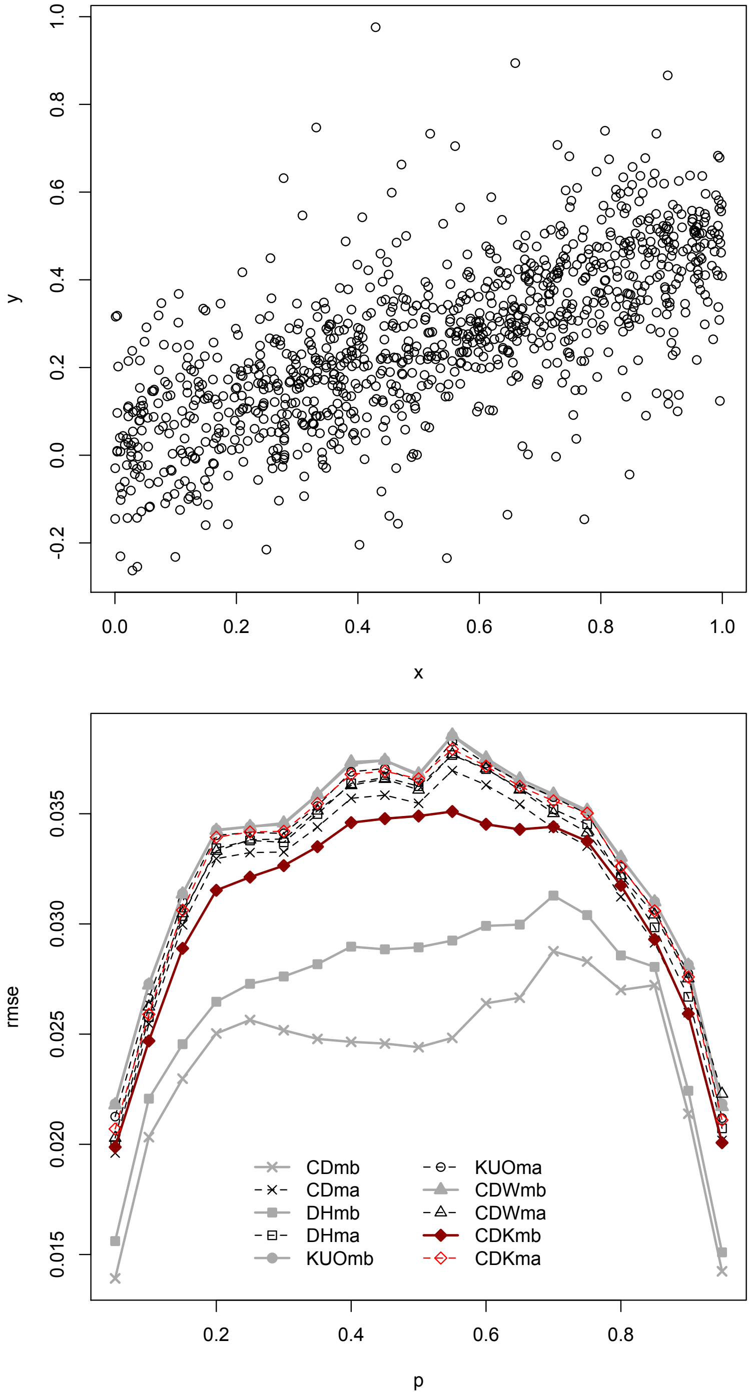

Fig. 1 LRIDE population and simulated rmse

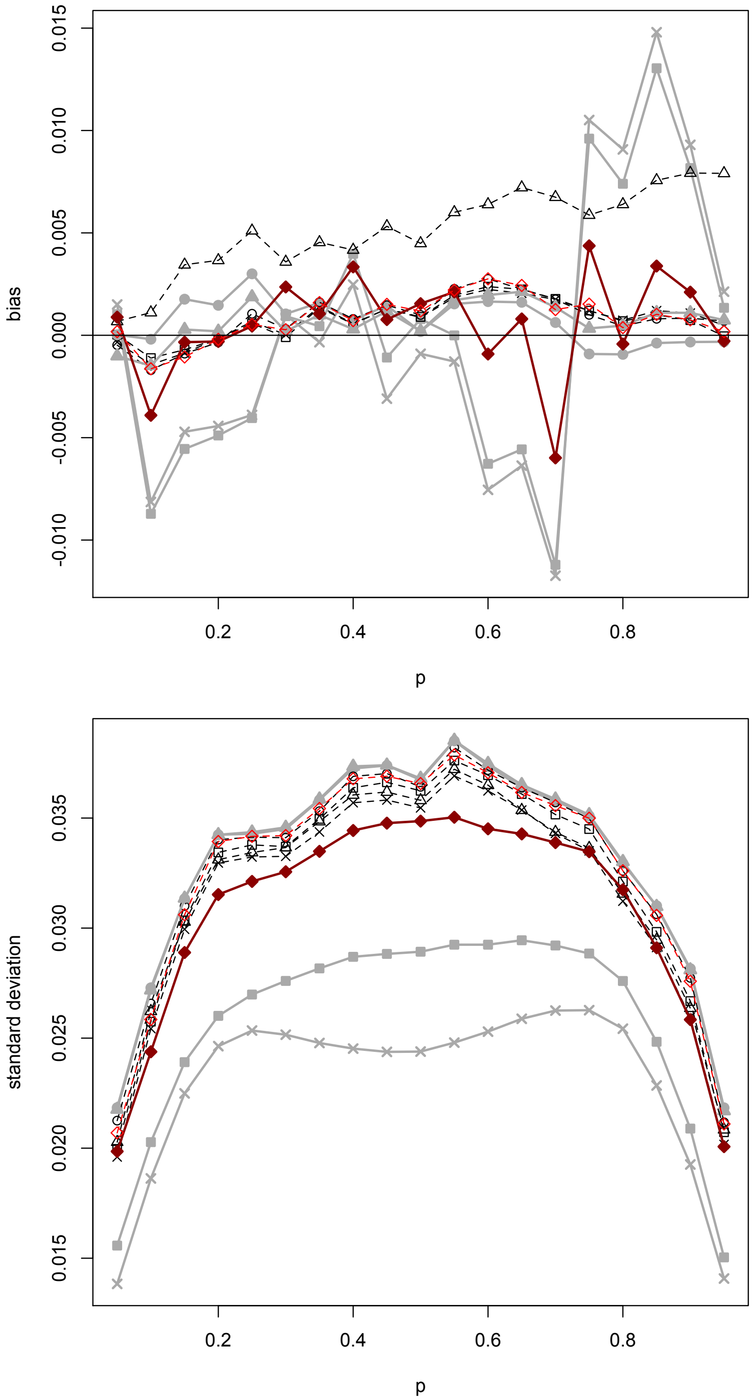

Fig. 2 LRIDE population: simulated bias and standard deviation

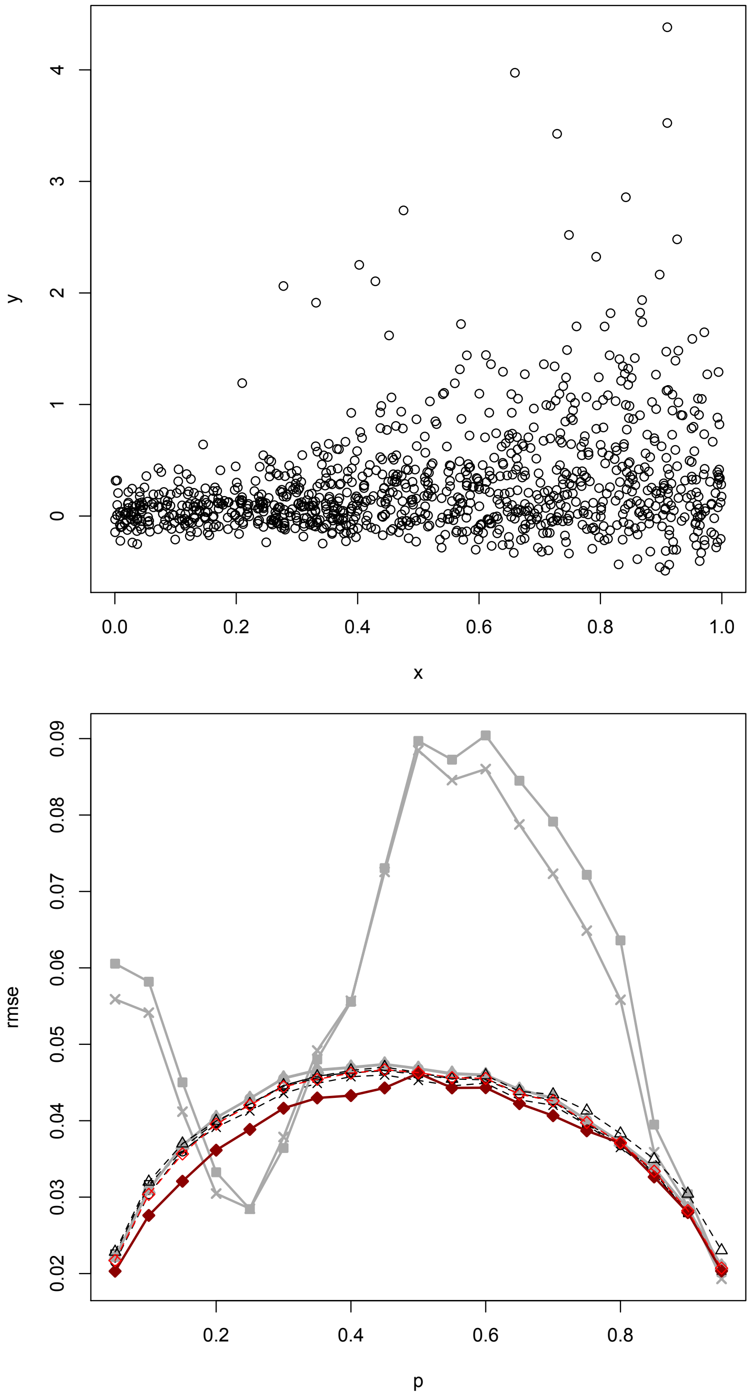

Fig. 3 LRNIDE population and simulated rmse

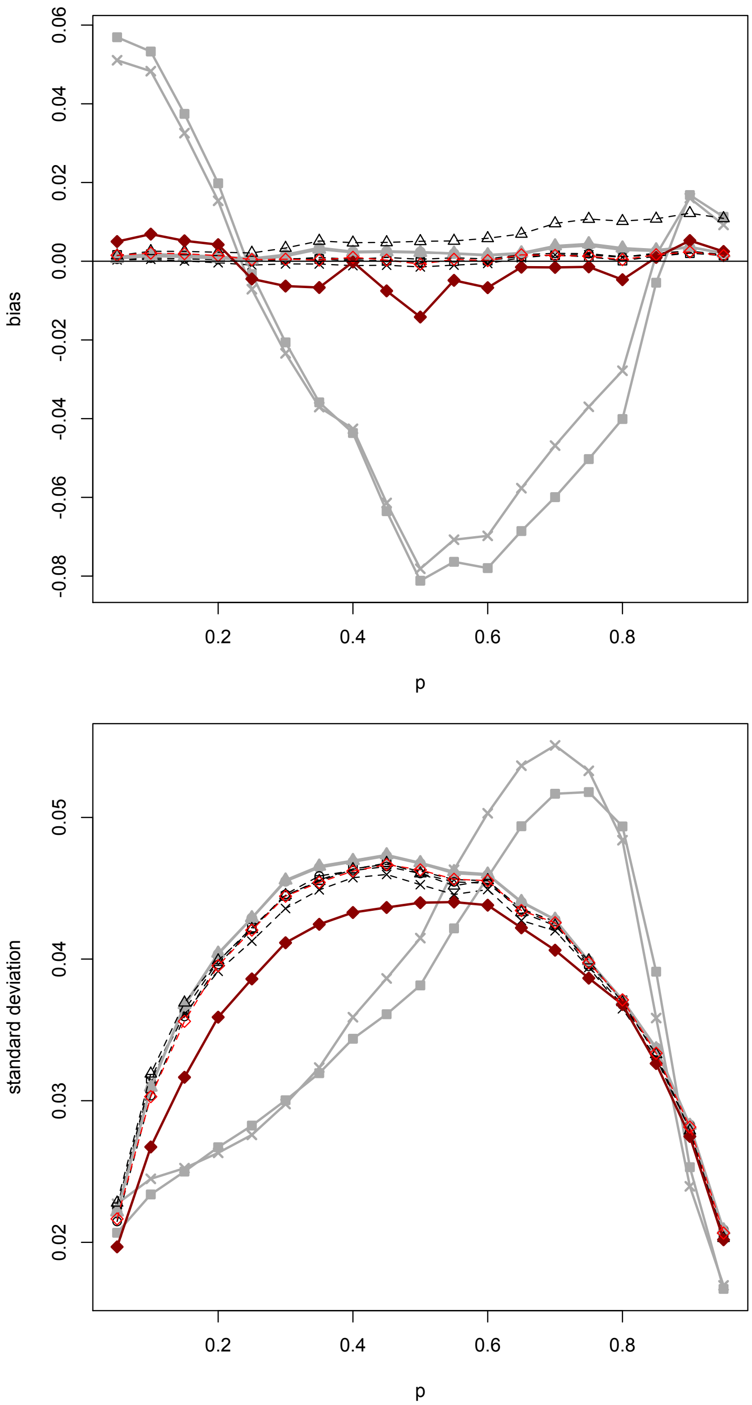

Fig. 4 LRNIDE population: simulated bias and standard deviation

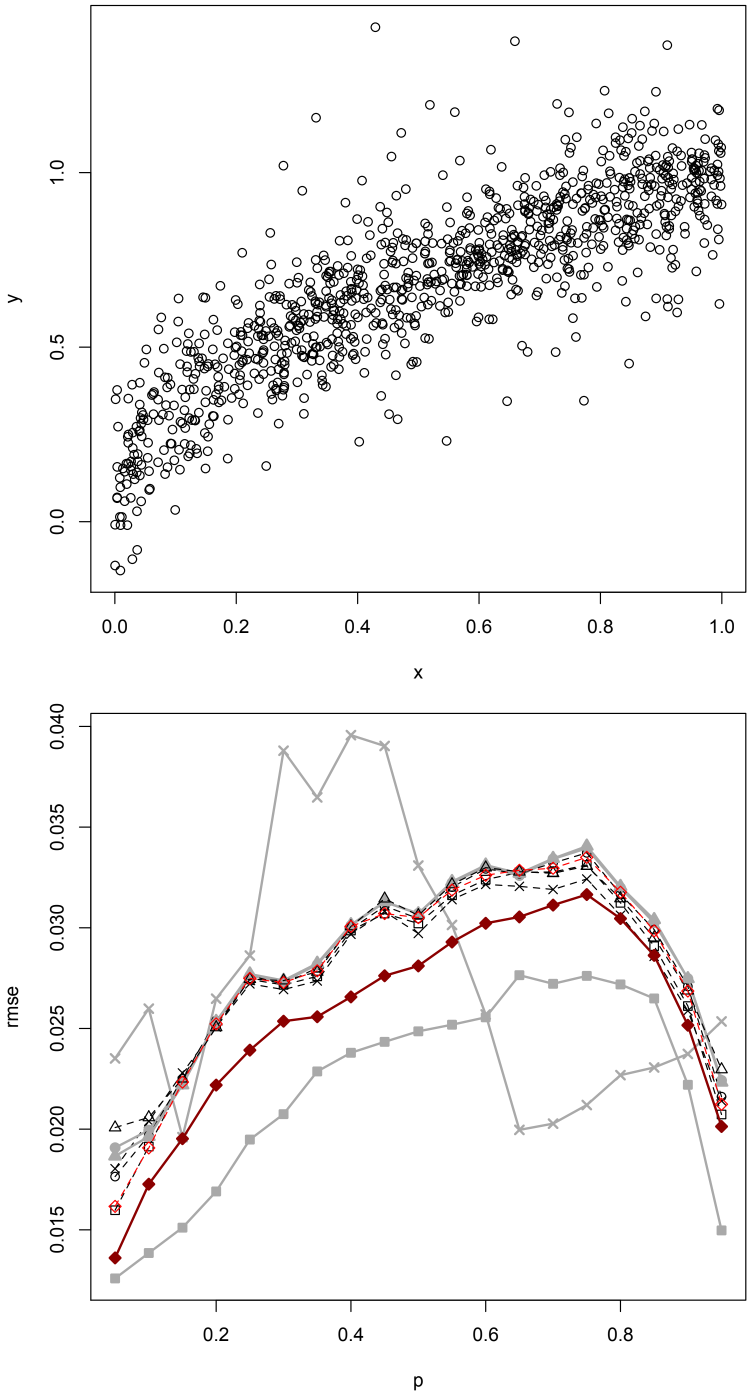

Fig. 5 NLRIDE population and simulated rmse

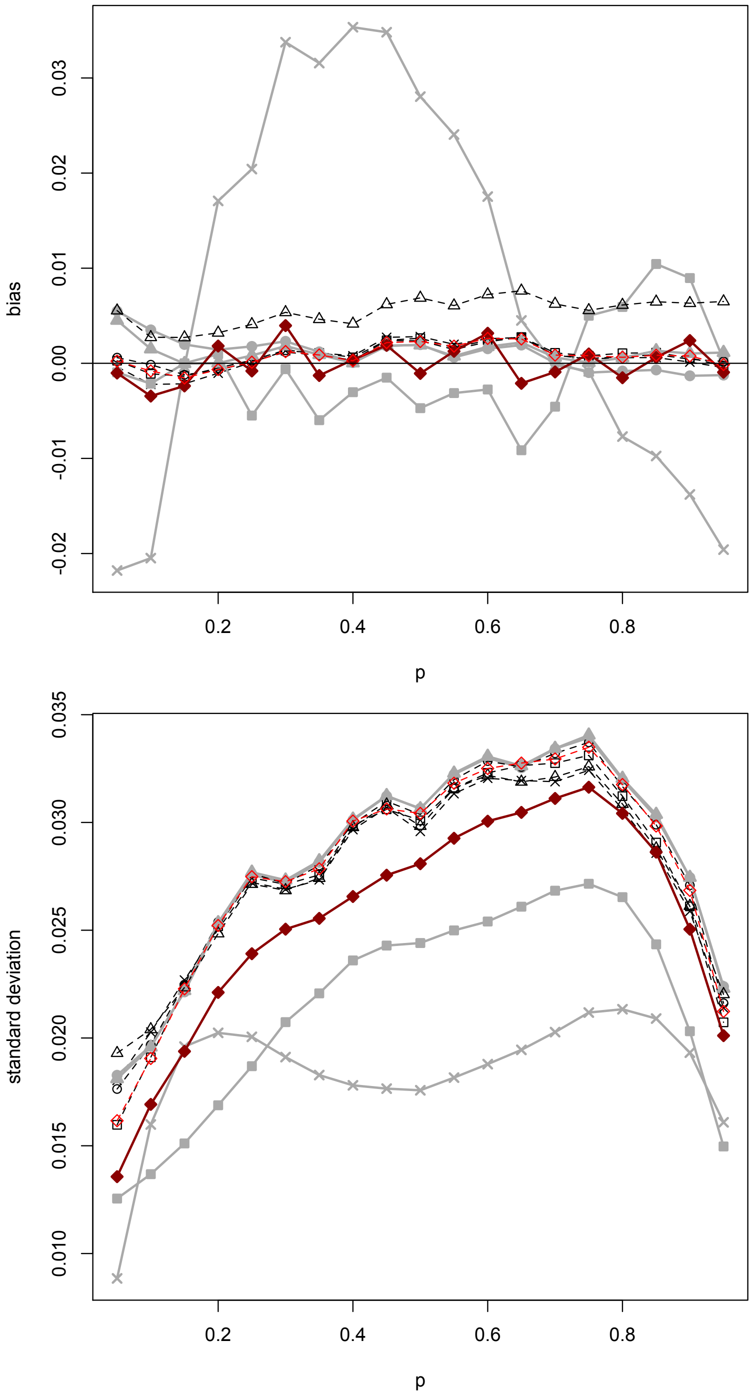

Fig. 6 NLRIDE population: simulated bias and standard deviation

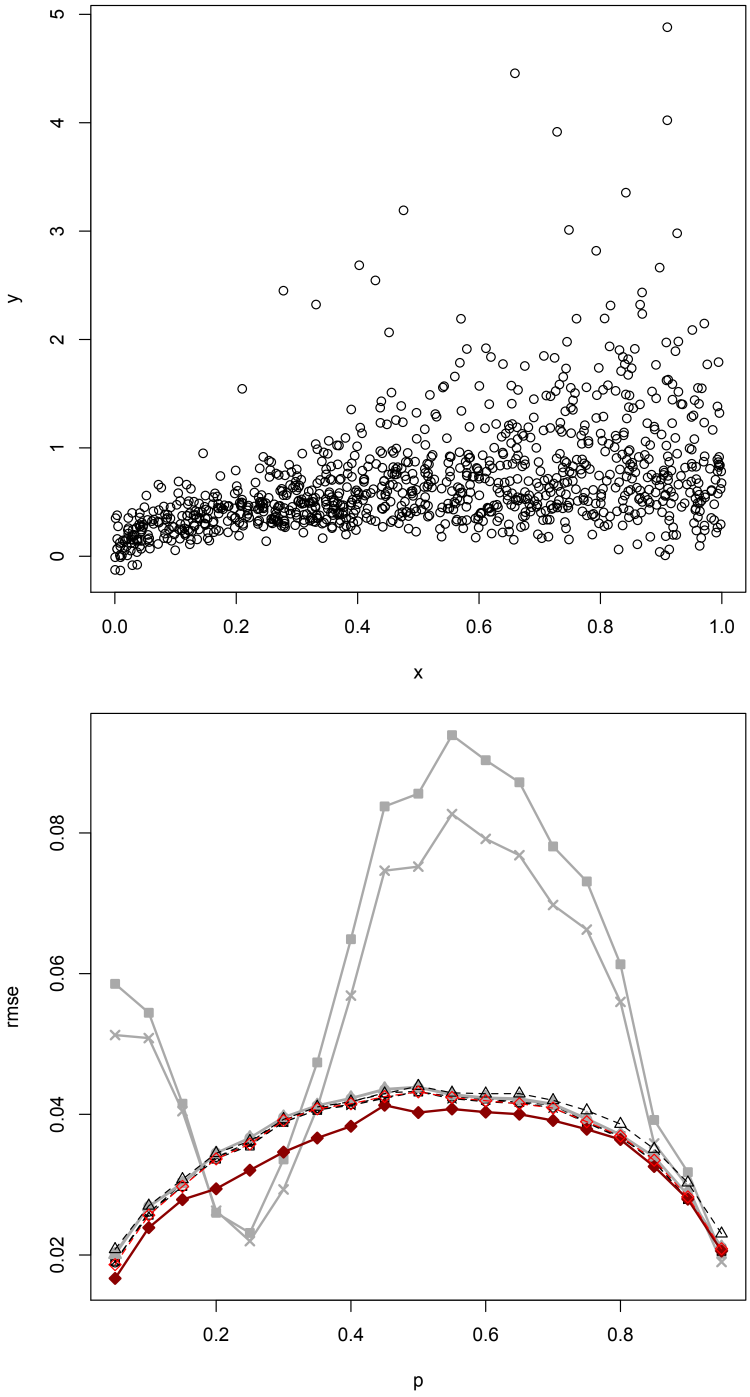

Fig. 7 NLRNIDE population and simulated rmse

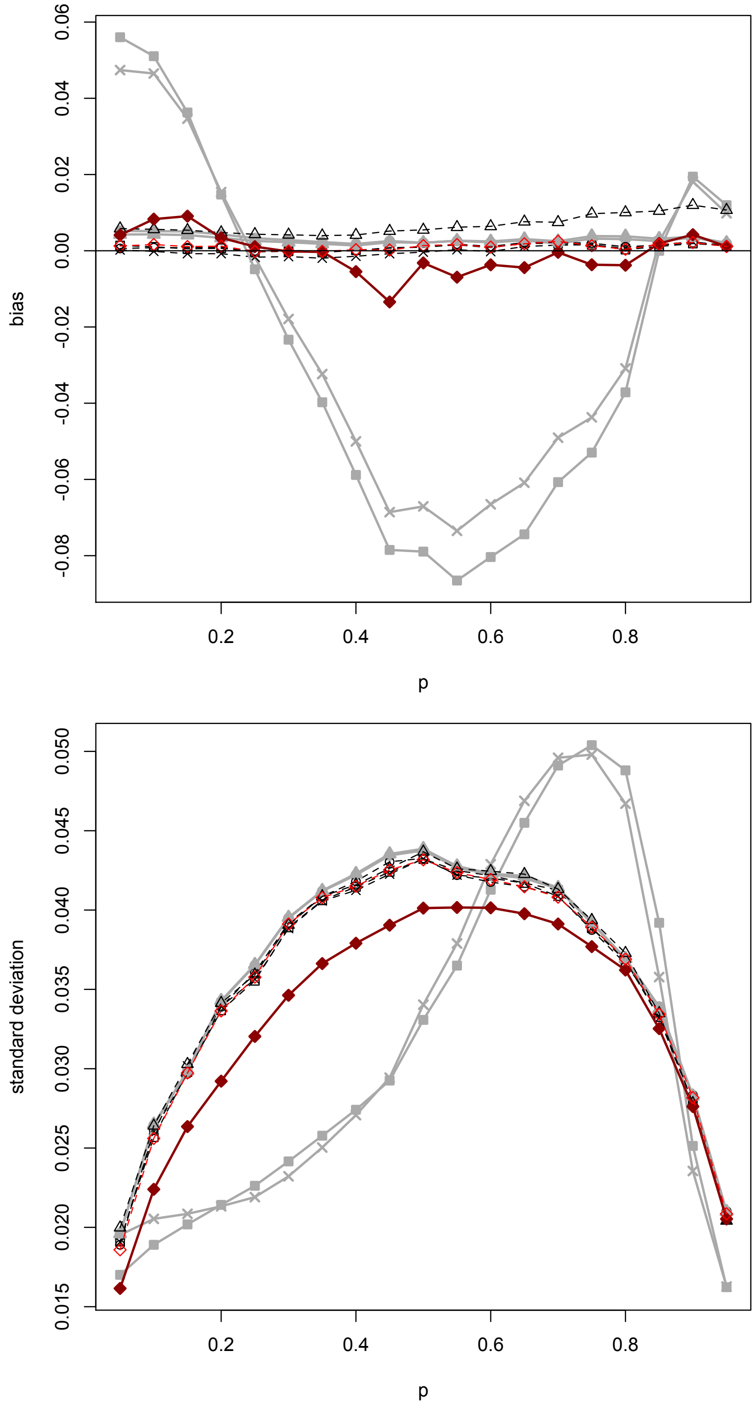

Fig. 8 NLRNIDE population: simulated bias and standard deviation