Small Area Estimation models with outliers in covariates

Monica Pratesi (monica.pratesi@unipi.it), Caterina Giusti, Stefano Marchetti, Nicola Salvati

Keywords: M-quantile, Robust estimation, Influence function.

1. INTRODUCTION

The objective of small area estimation (SAE) methods is to use survey data for estimating some characteristics, such as means, totals, quantiles of items of interest, in areas or domains where the sample size is not large enough to obtain reliable direct estimates. In the last years SAE methods have received a growing interest, since their use in the formulation of new policies and programs, poverty mapping and measurement of well-being indicators at detailed geographical level highly rely on these methods ([1]).

The most popular approach to model-based small area estimation are linear mixed models, that include random area effects to account for between area variations ([2]). Chambers and Tzavidis [3] proposed a new approach to SAE associated to M-quantile regression methods. M-quantile estimation is free of any distributional assumption, and is robust with respect to the presence of outliers and influential observations in the response variable.

The presence of outliers is a common feature in real data. Chambers [4] classifies outliers into two groups. Representative outliers are correctly measured sample values that are outlying relative to the rest of the sample data and for which there is no reason to believe that similar values do not exist in the non-sampled part of the survey population. Non-representative outliers are gross errors in the sample data, which have nothing to do with the values in the non-sampled part of the survey population. Either type of outlier can have a substantial impact on the estimates if ignored.

In the context of model-based estimation outliers can affect both the response and the auxiliary variables. The special case of outliers in the response variable resulting in the presence of outlying observations has been treated in the literature under both approaches to small area estimation, the mixed model and the M-quantile ones ([5], [6], [7]).

Sinha and Rao [5] suggest a way to extend their robust small area estimator to account for outliers in the auxiliary variables. In this paper we develop their estimator, and we propose a new M-quantile based small area estimator that account for outliers in the covariates. The two estimators are compared by means of model-based simulations both in the situation with representative and non-representative outliers.

2. METHODS

In a typical small area estimation problem we consider a population $U$ of size $N$ divided into $d$ non-overlapping subsets $U_i$ (domains of study or areas) of size $N_i$, $i = 1, \ldots, d$. With unit level models we commonly assume that a vector $x_{ij}$ of $p$ auxiliary variables is known without error for each unit $j$ belonging to area $i$, while the values of the variable of interest $y_{ij}$ are available only for a sample of population units in each area, $s_i \subset U_i$ of size $n_i \geq 0$. The set $r_i \subset U_i$ contains the $Ni - ni$ indices of the non-sample units in area $i$.

---

1 University of Pisa, Via C. Ridolfi 10, 56124 Pisa (PI), Italy.
We suppose that ζ% of the \( x_{ij} \) are representative-outliers or that ζ% of the sampled \( x_{ij} \) are non-representative-outliers, so that the covariates or the sampled covariates are \( x^*_{ij} = \{(1-\zeta%)x_{ij}, \zeta%(x_{ij} + \eta_{ij})\} \), where \( \eta_{ij} = C \).

We propose to deal with outliers in the covariates using a robust method that down-weights on the basis of extreme leverage values ([8]). We choose the trisquared redescending function to down-weight the values of the auxiliary variables - excluding the intercept. This function is defined as follows:

\[
w(t) = \left\{ 1 + \left( \frac{t}{k} \right)^3 \right\} l(\|t\|, k),
\]

where \( k \) is a tuning constant. For each observation we define the value \( z_j = (x_j^* - \mu)^T V(x_j^* - \mu) \) where \( \mu \) is a \( p \)-1 vector of 'robust' estimates of the centers of the \( p \)-1 auxiliary variables and \( V \) is a ‘robust’ estimate of the \((p-1) \times (p-1)\) covariance matrix of the auxiliary variables (without the intercept term). Defining \( u_j = (z_j / (p - 1))^{1/2} \), the weight to be used to down-weight the extreme \( p \)-1 values is

\[
w(t) = w(z_j^*) = \left\{ 1 + \left( \frac{u_j}{k} \right)^3 \right\} l(\|u_j\|, k).
\]

When the linear M-quantile model ([3], [9]) holds the proposed M-quantile small area estimator for the mean is as follows

\[
m^q_{\text{rob}} = \left( n + \sum_{j=1}^{n} w(x_j^*) \right) \left[ \sum_{j=1}^{n} y_j + \sum_{j=1}^{n} w(x_j^*) x_j^* \right] \hat{\beta}(q) + \frac{1}{n} \sum_{j=1}^{n} w(x_j^*) r_j
\]

where \( r_j = y_j - x_j^T \hat{\beta}(q) \) and \( \hat{\beta}(q) \) is obtained solving the following estimating equation with the iterative weight least square (IWLS) algorithm

\[
\sum_{j=1}^{n} \sum_{i=1}^{q} \left( y_j - x_j^T \hat{\beta}(q) \right) x_j^T = 0.
\]

Equation (4) takes into account both the influence function for outliers on the target variables (in the simulations we use the Huber proposal 2) and outliers on auxiliary variables. Indeed, for fixed \( q \) the estimator of the regression parameter, \( \hat{\beta}(q) \), is

\[
\hat{\beta}(q) = \left( X^T W(q) X \right)^{-1} X^T W(q) y,
\]

where \( W(q) \) is a diagonal matrix of order \( n \) which contains the final set of weights produced by the IWLS algorithm used to compute \( \hat{\beta}(q) \).

The other quantities in equations (3) and (4) are as in Tzavidis et al [9].

Note that the proposed estimator allows for the presence of representative and non-representative outliers in the covariates as well as outliers in the target variable. As a remark, the proposed estimator can account for outliers only for continuous variables. The same apply to our extension to the robust EBLUP estimator presented by Sinha and Rao [5].

The estimator proposed in equation (3) is based on the bias corrected version of the M-quantile estimator, see reference [9]. Using the estimated parameter \( \hat{\beta}(q) \) is
straightforward to extend the so-called naïve version of the M-quantile estimator (for the small area mean) to the case of outliers in the covariates. Under some settings the naïve M-quantile estimator can be more efficient than the bias corrected M-quantile estimator (as it is shown in reference [7]).

Using the same down-weight function we can obtain the robust version against outliers in the covariates of the robust (against outliers in the target) EBLUP $m^{\text{reblup}}$ originally proposed in [5]. This new version of the REBLUP is as follows

$$m^{\text{reblup rob}} = \left( \sum_{j=1}^{d} W(x_{ij}) \right)^{-1} \left[ \sum_{j=1}^{d} y_{ij} + \sum_{j=1}^{d} x_{ij}^T \hat{w} + \hat{u}_{i,w} \right],$$

(5)

where the parameters $w$ and $u_{i,w}$ are obtained solving the following estimating equation

$$\sigma X_i^T W V_i X_i = 0.$$

(6)

Matrix $W_i$ is a $n_i$ diagonal matrix with elements $w(x_{ij})$, $X_i$, $V_i$, and $U_i$ are as in [5].

3. **Results**

We use model-based Monte-Carlo simulations under several alternative scenarios to empirically evaluate the performance of the proposed small area robust estimators (3) and (5), comparing them with their original versions (not robust for outliers in the covariates), $m^{\text{eq}}$ and $m^{\text{reblup}}$, and with the classical EBLUP, $m^{\text{reblup}}$.

<table>
<thead>
<tr>
<th>Table 1. Results under Scenarios A and B, model-based simulations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\zeta = 1%$</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Setting A: outliers in the $y$, non-representative outliers in the $x$</td>
</tr>
<tr>
<td>$m^{\text{reblup}}$</td>
</tr>
<tr>
<td>$m^{\text{eq}}$</td>
</tr>
<tr>
<td>$m^{\text{reblup}}$</td>
</tr>
<tr>
<td>$m^{\text{reblup rob}}$</td>
</tr>
<tr>
<td>$m^{\text{reblup rob}}$</td>
</tr>
<tr>
<td>Setting B: outliers in the $y$, representative outliers in the $x$</td>
</tr>
<tr>
<td>$m^{\text{reblup}}$</td>
</tr>
<tr>
<td>$m^{\text{eq}}$</td>
</tr>
<tr>
<td>$m^{\text{reblup}}$</td>
</tr>
<tr>
<td>$m^{\text{reblup rob}}$</td>
</tr>
<tr>
<td>$m^{\text{reblup rob}}$</td>
</tr>
</tbody>
</table>

Under simulation scenarios A and B we consider the presence of outliers in the target variable and the presence of non-representative or representative outliers in the auxiliary variable, respectively. Population data for $d=30$ areas with $N=100$ are generated by using a unit level area random effects model with normally distributed random area effects and
unit level errors, $y_{ij} = 1 + 2x_{ij} + v_i + e_{ij}$, where the area random effects $v_i \sim N(0,3)$, but 10% come from the distribution $v_i \sim N(0,30)$, the unit level errors $e_{ij} \sim N(0,6)$, but 10% come from the distribution $e_{ij} \sim N(0,150)$. For the $x_{ij}$ we set $x_{ij} \sim N(5,1)$, with $\zeta$% of the sampled $x_{ij}$ affected by non-representative outliers (under scenario A) or with $\zeta$% of all the $x_{ij}$ affected by representative outliers (scenario B), as follows: $x_{ij}^* = \{1 - \zeta\% \} x_{ij}, \zeta\%(x_{ij} + 10\}$. The sample is obtained by selecting a within small areas random sample of $n_i = \{5, 10\}$ units from the corresponding population. To choose the value for the constant $k$ of the weight function $w$ we propose a cross validation criterion similar to the one proposed by Rudemo [10]. Results of the model-based simulations under scenarios A and B with $n_i = 5$ are shown in Table 1 in terms of percentage Relative Bias (RB%) and percentage Relative Root Mean Squared Error (RRMSE%).

4. CONCLUSIONS

The results of the model-based simulations are encouraging: estimators $\bar{m}_{q}^{\text{rob}}$ and $\bar{m}_{q}^{\text{reblup-rob}}$ are able to down-weight the effect of the presence of non-representative and representative outliers in the auxiliary variable, both in bias and variability. In further extensions we will develop a naïve version of the proposed M-quantile robust estimators as well as a bootstrap estimators of the mean squared error for the presented estimators.

REFERENCES


