

Design consistent small area estimators based on M-quantile regression

Fabrizi Enrico, Pratesi Monica, Salvati Nicola, Tzavidis Nikos

Abstract Small area estimators based on M-quantile regression have been recently introduced by [4]. The aim is to obtain reliable small area estimates robust with respect to the presence of outliers and without the demanding distributional assumptions typical of the predictors based on linear mixed models. These M-quantile predictors do not make use of the unit level survey weights and, as a result, they are not design consistent as the area sample sizes become large, unless the sampling design is self-weighting within areas. In this paper, exploiting a GREG-like representation of the bias-adjusted M-quantile predictors, we introduce design consistent small area estimators based on quantile regression that use unit-level sampling weights. These estimators are compared with pseudo-EBLUP predictors, and a bootstrap algorithm for the estimation of their design-based variance is introduced.

Key words: Finite populations, sampling weights, robust estimation, small area estimation, bootstrap procedure

Fabrizi Enrico
DISES, Università Cattolica di Piacenza e-mail: enrico.fabrizi@unicatt.it

Pratesi Monica
Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa e-mail: m.pratesi@ec.unipi.it

Salvati Nicola
Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa e-mail: salvati@ec.unipi.it

Tzavidis Nikos
Centre of Census and Survey Research, University of Manchester e-mail: nikos.tzavidis@manchester.ac.uk

1 Introduction

In statistical inference about a finite population, estimates of population descriptive quantities for a target variable y are usually needed for the population as a whole and for different collections of sub-populations (domains or areas). A small area estimation problem arises when the available samples are not large enough to allow for reliable estimation using standard design-based estimators for all or most of the domains (areas) being considered. A large and growing literature is devoted to this subject; see [11] for a general introduction to the topic.

All small area estimation methods are based on the availability of population-level auxiliary information to improve the precision of estimation process. They have their own specificity in the way of linking auxiliary information and target variable and in the properties of the obtained small area predictors.

Unit-level models, including nested error regression [1], multi-level [10] and, more recently, quantile regression models [4] are used to obtain efficient model-based predictors of small area means and other parameters. In particular we will focus on this latter class of models; they are based on the assumption of a quantile-specific linear relationship between y and x , and, differently from nested error regression and multilevel models are free of distributional assumptions. Moreover the recourse to M-estimation of quantile regression coefficients as in [4], protects from presence of outliers and influential observations and do not require explicit specification of the random part of the model.

Model-based estimators typically do not make use of the unit level survey weights; as a result the estimators are not design consistent as the area sample sizes become large, unless the sampling design is self-weighting within areas, but this is seldom the case in practice. Several methods have been proposed to make use of survey weights in small area estimation based on linear mixed and generalized linear mixed models ([8], [12], [6]). This paper proposes a design consistent estimator based on M-quantile regression.

In section 2 the estimator is introduced and its properties explored. In section 3 we introduce a simulation exercise comparing weighted and non-weighted M-quantile based small area estimators along psuedo-EBLUP (design consistent) and EBLUP predictors associated to linear mixed models.

2 A design-consistent small area estimator based on M-quantile regression

Let's suppose that a population U of size N is divided into m non overlapping subsets (domains of study or areas) of size N_i , $i = 1, \dots, m$. We are interested in a target variable y and more specifically in estimating the area level means $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$. Suppose that a sample is drawn according to some, possibly complex, sampling design from the population, so that area-specific samples of size $n_i \geq 0$ are available

for each area. Values of y_{ij} and p -vector of auxiliary variables $\mathbf{x}_{ij}$ are known only for sampled values; moreover we assume that area means of auxiliary variables $\bar{\mathbf{X}}_i$ are accurately known.

A recently proposed approach to small area estimation is based on the use of M-quantile models (see [4]). Since much of the development in this paper is based on the application of linear quantile/M-quantile regression, we now give a brief definition of these concepts. Ordinary linear regression is based on the idea of modelling the expected value of the dependent variable as a function of the regressors; that is, in our notation, on the assumption that $E(y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \beta$. In quantile regression it is the q -th quantile that is assumed to be a linear function of the auxiliary information, i.e. $Q_q(y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \beta(q)$, $q \in (0, 1)$. This means that a distinct (hyper)plane is fitted to the data for each $q \in (0, 1)$ according to quantile-specific regression coefficients $\beta(q)$. See [7] for a general introduction to quantile regression. The vector of $\beta(q)$ may be determined according to some minimization criteria such as least absolute deviations. [2] introduced the application of robust M-estimation to quantile regression. M-quantile regression provides a ‘quantile-like’ generalization of regression based on influence functions. For specified q and influence function ψ , estimate of the vector of the regression parameters $\beta_\psi(q)$ may then be obtained by solving the following normal equations:

$$\sum_{i=1}^m \sum_{j=1}^{n_i} \psi_q(y_{ij} - \mathbf{x}_{ij}^T \beta_\psi(q)) \mathbf{x}_{ij} = \mathbf{0},$$

in $\beta_\psi(q)$, where $\psi_q(r) = 2\psi(s^{-1}r)\{qI(r > 0) + (1 - q)I(r \leq 0)\}$. The influence function ψ may for instance chosen as in Huber proposal 2 (see [5]), i.e. $\psi(r) = rI(-c \leq r \leq c) + c\text{sign}(r)$ as we will do in the application of next section; in this case c is a tuning constant assumed to be bounded away from 0. The quantity s in the definition of $\psi_q(r)$ is a robust estimate of the scale of the data such as the mean absolute deviation $s = \text{med}|r|/0.6745$. For specified q an estimate $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ is then obtained via iterative re-weighted least squares.

Following [4] an alternative to random effects for characterizing the variability across the population not accounted for by the regressors, is to use the M-quantile coefficients of the population units. For unit j in area i with values y_{ij} and $\mathbf{x}_{ij}$, this coefficient is the value θ_{ij} such that $Q_{\theta_{ij}}(y_{ij}|\mathbf{x}_{ij}; \psi) = y_{ij}$. [4] observe that if a hierarchical structure does explain part of the variability in the population data, units within areas defined by this hierarchy are expected to have similar M-quantile coefficients. When the conditional M-quantiles are assumed to follow a linear model, with $\beta_\psi(q)$ a sufficiently smooth function of q , the following predictor of $\bar{Y}_i$ may be proposed:

$$\hat{Y}_i^{MQ} = N_i^{-1} \left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \mathbf{x}_{ij}^T \hat{\beta}_\psi(\hat{\theta}_i) + \frac{N_i - n_i}{n_i} \sum_{j \in s_i} \{y_{ij} - \mathbf{x}_{ij}^T \hat{\beta}_\psi(\hat{\theta}_i)\} \right], \quad (1)$$

where $\hat{\theta}_i$ is an estimate of the average value of the M-quantile coefficients of the units in area i . These unit level coefficients are estimated by solving $\hat{Q}_{q_j}(y_{ij}|\mathbf{x}_{ij}; \psi) =$

y_{ij} for q_j with $\hat{Q}_q$ denoting the estimated value of $Q_q(y_{ij}|\mathbf{x}_{ij}; \psi)$ at q . When there is no sample in area i then $\hat{\theta}_i = 0.5$.

We may refer to (1) as the bias adjusted M-quantile predictor of $\bar{Y}_i$ that is derived as the mean functional of the [3] estimator of the distribution function. It may easily be shown that (1) has an interesting GREG-like representation:

$$\hat{Y}_i^{MQ} = n_i^{-1} \sum_{j \in s_i} y_{ij} + \left\{ N_i^{-1} \sum_{j \in U_i} \mathbf{x}_{ij}^T - n_i^{-1} \sum_{j \in s_i} \mathbf{x}_{ij}^T \right\} \hat{\beta}_\psi(\hat{\theta}_i). \quad (2)$$

Put in the form (2), $\hat{Y}_i^{MQ}$ may easily be modified to make use of unit level survey weights; that is

$$\hat{Y}_{iw}^{MQ} = \bar{y}_{iw} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{iw})^T \hat{\beta}_\psi(\hat{\theta}_i) = \sum_{j \in s_i} w_{ij} y_{ij} + \left(\sum_{j \in U_i} \mathbf{x}_{ij}^T - \sum_{j \in s_i} \mathbf{x}_{ij}^T w_{ij} \right) \hat{\beta}_\psi(\hat{\theta}_i), \quad (3)$$

where $w_{ij} = \tilde{w}_{ij} / \sum_{j \in s_i} \tilde{w}_{ij}$ and $\tilde{w}_{ij} = \pi_{ij}^{-1}$ are the inverse of the inclusion probabilities. If we assume $\beta_\psi(\theta_i)$ known, it is straightforward to note that $\hat{Y}_{iw}^{MQ}$ is design unbiased and consistent under ordinary, mild conditions on the sampling design. For a given quantile q , a weighted estimator $\hat{\beta}_{\psi w}(q)$ may be obtained as the solution of the following system of estimating equations:

$$\sum_{i=1}^m \sum_{j \in s_i} \tilde{w}_{ij} \psi_q(y_{ij} - \mathbf{x}_{ij}^T \beta(q)) \mathbf{x}_{ij} = 0, \quad (4)$$

using re-weighted least squares. Note that $\hat{\beta}_{\psi w}(\theta_i)$ is a design consistent estimator of $B_{\psi w}(\theta_i)$ defined as the solution of the following system of ‘Census’ equations:

$$\sum_{i=1}^m \sum_{j \in U_i} \psi_q(y_{ij} - \mathbf{x}_{ij}^T \beta(q)) \mathbf{x}_{ij} = 0. \quad (5)$$

Design consistency of $\hat{Y}_{iw}^{MQ}$ follows regardless of the chosen θ_i . In practice, a feasible $\hat{Y}_{iw}^{MQ}$ will be obtained replacing q with an estimated $\hat{\theta}_i$ along the lines illustrated above.

For fixed $\hat{\theta}_i$ a variance estimator based on Taylor linearization for area i is given by

$$\hat{V}(\hat{Y}_{iw}^{MQ}) = \hat{N}_i^{-2} \sum_{j \in s_i} \sum_{k \in s_i} \frac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{\pi_{ijk}} \frac{(y_{ij} - \mathbf{x}_{ij}^T \beta(\hat{\theta}_i))}{\pi_{ij}} \frac{(y_{jk} - \mathbf{x}_{jk}^T \beta(\hat{\theta}_i))}{\pi_{ik}}, \quad (6)$$

where $\hat{N}_i = \sum_{j \in s_i} w_{ij}$. The estimator is a first order approximation because (6) does not take into account both the variability due to the estimation of the value $\hat{\theta}_i$ and that associated to the estimation of $\hat{\beta}_{\psi w}(q)$. Since the M-quantile estimator is an approximately unbiased estimator of the small area mean, the squared bias will not impact significantly on the estimator’s mean square error $MSE(\hat{Y}_{iw}^{MQ})$. An estimator

of the MSE taking into account all sources of variability, can be found implementing a bootstrap procedure ([9], chapter 11). The bootstrap procedure is briefly described as follows:

- Using the sample data to build an artificial population U^* by using the sample weights w_{ij} .
- Draw T independent samples s^* from U^* by using a design identical to the one used to draw s from U .
- For each bootstrap sample compute the replication $\hat{Y}_{itw}^{MQ*}$ of $\hat{Y}_{itw}^{MQ}$ for $i = 1, \dots, m$ and $t = 1, \dots, T$.

The bootstrap estimator of the MSE of $\hat{Y}_{itw}^{MQ}$ for each small area can be computed as

$$\widehat{MSE}_i = T^{-1} \sum_{t=1}^T \left(\hat{Y}_{itw}^{MQ*} - \hat{Y}_{itw}^{MQ} \right)^2.$$

3 A simulation exercise

In this section design-based simulation experiments are carried out to evaluate the performance of the $\hat{Y}_{itw}^{MQ}$ estimators in the context of real population and realistic sampling methods. The population underpinning the simulation, denoted MU284, is taken from [9]. The data consists of $N = 284$ municipalities in Sweden. The variable of interest is the revenues from 1985 municipal taxation ($y = RMT85$) and it follows a highly skewed distribution. The covariate is the real estate values according to 1984 assessments for each municipality ($x = REV84$). The municipalities making up the population are grouped into 8 regions that are defined, in our simulations, as the small area of interest. Samples of municipalities are drawn under the Poisson sampling scheme with three different methods to construct the selection probabilities:

- Scenario 1: $\pi_j \propto 0.2x_j + 0.05$ where $x_j = REV84_j$ is highly correlated with the study variable ($\rho > 0.9$);
- Scenario 2: $\pi_j \propto SS82_j$ where $SS82$ is the number of Social-Democratic seats in municipal council and it has a medium correlation with the study variable ($\rho \approx 0.3$);
- Scenario 3: $\pi_j \propto z$ where z are generated as independently and identically distributed from a lognormal distribution with a mean of 1 and a standard deviation of 3 on the log scale and it is not correlated with the study variable.

Each scenario is independently simulated 10000 times. The predictor (3) is compared to the $\hat{Y}_i^{MQ}$, the P-EBLUP of [12] and the EBLUP estimator of [1]. The performance of these estimators across the different areas and simulations is assessed by computing the mean and the median values of their area specific relative bias (Rbias) and relative root mean squared error (RRMSE). Table 1 sets out these mean and median values for the different simulation scenarios and different estimators.

Mean area revenues are estimated with less bias under the MQ approach, even when MQ estimator is crude as not modified by the sampling weights. This is probably due to the presence of outliers in the data and to the skewness of the distribution of the study variable. Under varying probabilities sampling designs, the result is confirmed, but the major effect in reducing bias is achieved when the probabilities are highly correlated with the study variable and they are included in the expression of the estimator. In Scenario 1 the relative bias of $\hat{Y}_w^{MQ}$ is about one half of the P-EBLUP bias, and in Scenario 2, even if the correlation is reduced, the $\hat{Y}_w^{MQ}$ estimator has the minimum value of bias (mean RBias 0.39) of the whole result set of the simulation. As concerns the variability, MQ and EBLUP approach to estimate the mean give competing results when the estimators are not weighted, although it seems on average that EBLUP is slightly more accurate in this simulation set. However, when introducing highly correlated weights in Scenario 1, $\hat{Y}_w^{MQ}$ is less variable than P-EBLUP. The gain in precision is not confirmed when we relax the correlation between the probabilities and the study variable (Scenario 2) or we generate independent weights (Scenario 3).

Table 1 Mean and median values of the Rbias and RRMSE for the three scenarios and the small area predictors.

		$\hat{Y}_i^{MQ}$	$\hat{Y}_{iw}^{MQ}$	EBLUP	P-EBLUP
Scenario	Index	Rbias			
Scenario 1	Median	-19.96	-0.97	-48.33	-3.80
	Mean	-11.17	-1.60	-48.19	2.11
Scenario 2	Median	1.06	0.39	-6.13	-1.88
	Mean	2.37	0.62	-1.90	1.42
Scenario 3	Median	-7.73	-6.00	-10.83	-10.82
	Mean	-4.98	-4.39	-7.53	-7.08
		RRMSE			
Scenario 1	Median	34.74	14.64	54.76	16.32
	Mean	51.63	17.22	62.22	22.38
Scenario 2	Median	25.32	22.14	22.90	18.87
	Mean	29.69	24.58	25.79	23.53
Scenario 3	Median	22.59	27.84	20.10	23.43
	Mean	21.73	27.13	18.96	23.65

In the simulation we have also investigated the performance of the bootstrap procedure of MSE estimation. The behaviour of the 'true' root MSE estimated from MC replications and its estimator for each area is shown in Figure 1. It can be seen that the bootstrap MSE estimator for the MQ tracks the irregular profile of the area-specific empirical MSE very well.

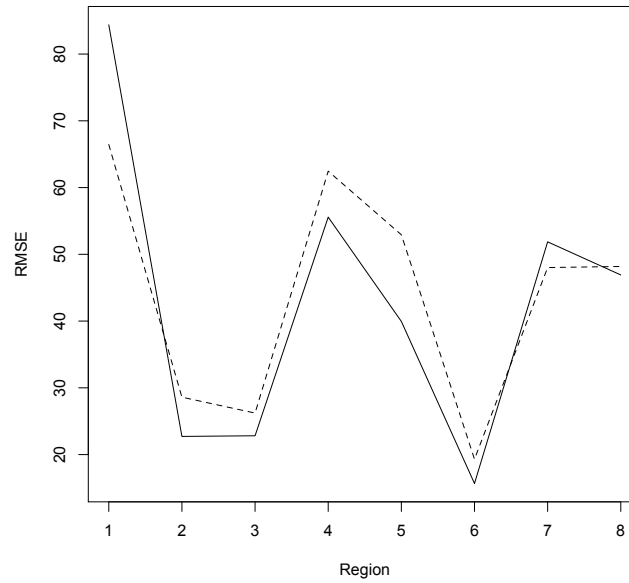


Fig. 1 Region-specific values of actual design-based RMSE (solid line) and average estimated RMSE (dashed line). The estimates of RMSE are obtained by bootstrap procedure.

Acknowledgements This work is financially supported by the European Project SAMPLE “Small Area Methods for Poverty and Living Condition Estimates”, European Commission 7th FP - www.sample-project.eu, and by the Research Project PRIN2007 “Inference in the presence of imperfect auxiliary information: sampling from elusive populations and estimation for unplanned domains”.

References

1. Battese G.E., Harter R.M., Fuller W.A.: An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*. **83**, 28–36 (1988).
2. Breckling J., Chambers R.: M-quantiles. *Biometrika*. **75**, 761–771 (1988).
3. Chambers R., Dunstan R.: Estimating distribution functions from survey data. *Biometrika*. **73**, 597–604 (1986)
4. Chambers R., Tzavidis N.: M-quantile Models for Small Area Estimation. *Biometrika*. **93**, 255–268 (2006)
5. Huber P.J.: *Robust statistics*. Wiley, New York (1981)
6. Jiang J., Lahiri P.: Estimation of finite population domain means: a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*. **101**, 301–311 (2006).
7. Koenker R., Bassett G.: Regression quantiles. *Econometrica*. **101**, 33–50 (1978).

8. Kott P.: Robust small domain estimation using random effects modelling. *Survey Methodology*. **15**, 1–12 (1989)
9. Särndal C.E., Swensson B., Wretman J.: *Model Assisted Survey Sampling*. Springer-Verlag, New York (1992)
10. Rao J.N.K.: Some recent advances in model-based small area estimation. *Survey Methodology*. **25**, 175–186 (1999)
11. Rao, J.N.K.: *Small Area Estimation*. Wiley, New York (2003)
12. You Y., Rao J.N.K.: A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*. **30**, 431–439 (2002)