M-quantile models for Small Area Estimation

N. Salvati¹, E. Fabrizi²

¹DEM, Pisa
²Piacenza
Introduction to Small Area Estimation

- Problem: demand from official and private institutions of statistical data referred to a given population of interest
- Possible solutions:
  - Census
  - Sample survey

Sample surveys have been recognized as cost-effectiveness means of obtaining information on wide-ranging topics of interest at frequent interval over time
Introduction to Small Area Estimation (Cont´d)

- Population of interest (or target population): population for which the survey is designed
  → *direct estimators* should be reliable for the target population

- Domain: sub-population of the population of interest, they could be planned or not in the survey design
  - Geographic areas (e.g. Regions, Provinces, Municipalities, Health Service Area)
  - Socio-demographic groups (e.g. Sex, Age, Race within a large geographic area)
  - Other sub-populations (e.g. the set of firms belonging to a industry subdivision)

  → we don’t know the reliability of *direct estimators* for the domains that have not been planned in the survey design
Often *direct estimators* are not reliable for some domains of interest.

In these cases we have two choices:

- oversampling over that domains
- applying statistical techniques that allow for reliable estimates in that domains

**Small Domain or Small Area**

Geographical area or domain where direct estimators do not reach a minimum level of precision.

**Small Area Estimator (SAE)**

An estimator created to obtain reliable estimate in a Small Area.
Notation & Assumptions

- Individual level covariates $x$
- Area level covariates $z$
- Area random effect $v$
- Population / Sample / Non-sample $U/s/r$

Assumption 1
Linear relationship between $y$ and $x$, $z$

Assumption 2
All small areas sampled

Assumption 3
Small area means of $x$ and $z$ are known

Working Model \( y = X\beta + Zv + e \)

Assumptions \( u \sim N(0, \Sigma_v), \ e \sim N(0, \Sigma_e) \)

Empirical Best Linear Unbiased Predictor of \( \bar{y}_i \) is

\[
\hat{y}_i = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (x_{ij}^T \hat{\beta} + z_{ij}^T \hat{v}_i) \right\} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij} \right\}
\]
Random Intercepts Modelling of Grouped Data
The M-quantile approach to SAE

- The M-quantile approach to small area estimation has been proposed by Chambers and Tzavidis (2006).
- This method is based on the M-quantile regression model and it is an alternative to the methods that are based on the mixed effect models.
- The M-quantile regression is a generalized robust model to handle the tail of a conditional distribution.
- The estimators we present here are based on the M-quantile linear model with a Huber proposal II loss function.
An overview of M-quantile models

- Traditionally, with regression models we model the expectation of the conditional distribution $f(y|X)$.
- An approach to outlier robust regression analysis is M-regression that is based on the use influence functions for controlling the effect of outliers.
- M-regression controls the effect of outliers by treating a residual whose magnitude is greater than a given cutoff, $c$, as if its magnitude equals $c$.
- With M-regression we model the median of the conditional distribution $f(y|X)$.
- A more complete picture is offered by modeling not only measures of central tendency of $f(y|X)$ but also other quantiles. This takes us to the idea of M-quantile regression.
M-quantile models more formally

- The M-quantile model for the $q$th quantile of $f(y|X)$ is

$$Q_q(y|X) = X\beta_{\psi}(q)$$

- Estimates of $\beta_{\psi}(q)$’s are obtained via Iterative Weighted Least Squares (IWLS) by solving

$$\sum \psi_q \left\{ y_i - x_i^T \beta_{\psi}(q) \right\} x_i = 0$$

$$\hat{\beta}_{\psi}(q) = (X^T WX)^{-1} X^T Wy$$

- $\psi_q$ denotes an influence function that controls the effect of outliers
- $W$ is an $n$ by $n$ diagonal weighting matrix that depends on both the influence function and the quantile one models
The influence function $\psi_q$ is

$$
\psi_q\{u\} = \begin{cases} 
2q\psi\{u\} & u \geq 0 \\
2(1 - q)\psi\{u\} & u < 0
\end{cases}
$$

The influence function $\psi$ is Huber Proposal II

$$
\psi\{u\} = ul(-c < u < c) + csgn(u)
$$
Why Regression M-quantiles?

- Standard regression quantile fitting algorithms are based on linear programming methods and do not necessarily guarantee convergence and a unique solution. In contrast, the simple IRLS algorithm used to fit a regression M-quantile is guaranteed to converge to a unique solution for a continuous monotone influence function.
- Bianchi and Salvati (2014) proposed also an analytical estimator for the standard errors on regression M-quantile coefficients.
- M-quantile models allow for more flexibility in modelling. For example, the tuning constant $c$ defining the Huber influence function can be used to trade outlier robustness for efficiency.
Quantile and Expectile Regression

- M-quantile regression includes, as special cases, quantile regression (Koenker & Bassett, 1978) and expectile regression (Newey & Powell, 1987)

- Quantile regression is obtained by allowing the cutoff $c$ in M-quantile regression to approach 0. In this case the weight given to a residual depends only on its sign and not its magnitude

- Expectile regression, on the other hand, corresponds to allowing $c$ to be infinitely large, so the weight given to a residual depends on the magnitude of the residual
Quantile and Expectile Regression (Cont’d)
Using M-quantile Models in Small Area Estimation

Reference Chambers and Tzavidis (2006)

- Each sample unit $j \in i$ will lie on one and only one regression M-quantile line. The $q$-index of this line is the M-quantile coefficient $q_{ij}$ of sample unit $j$.
- The M-quantile coefficient $q_i$ for area $i$ is a suitable average of the M-quantile coefficients $q_{ij}$ in area $i$.
- Naive M-quantile estimate of the area $i$ mean is

$$\hat{y}_i^{MQ} = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} x_{ik}^T \hat{\beta}_q \right) = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} \hat{y}_{ik}^{MQ} \right)$$
Regression M-quantile Modelling of Grouped Data
A Research Update (1)

- Extended version of the M-quantile model for the estimation of the small area distribution function using a non-parametric specification of the conditional MQ of the response variable given the covariates (Pratesi, Ranalli and Salvati, 2008)

\[
\hat{y}_{i}^{NPMQ} = N_{i}^{-1}\left(\sum_{j \in s_{i}} y_{ij} + \sum_{k \in r_{i}} \left\{ x_{ik}^{T} \hat{\beta}_{qi} + z_{ik}^{T} \hat{\gamma}_{qi} \right\} \right)
\]

\[
= N_{i}^{-1}\left(\sum_{j \in s_{i}} y_{ij} + \sum_{k \in r_{i}} \hat{y}_{ik}^{NPMQ}\right)
\]

- Robust prediction of small area means and distributions (Tzavidis, Marchetti and Chambers, 2010)
Spatial version of the M-quantile models for small area estimation
(Salvati, Tzavidis, Pratesi and Chambers, 2012)

\[ \hat{y}_i^{MQGWR} = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} x_{ik}^T \hat{\beta}_i (g_{ik}) \right) \]

\[ = N_i^{-1} \left( \sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} \hat{y}_{ik}^{MQGWR} \right) \]
A Research Update (3)

- Constrained M-quantile small area estimators for benchmarking and for the correction of the under/over-shrinkage of small area estimators (Fabrizi, Salvati, Pratesi, 2012). More specifically, given a set of predictors \( \{ \hat{y}_i^{MQ} \}_{1 \leq i \leq m} \) of the small area means, we look for a new set of estimators \( \{ t_i^{MQ} \}_{1 \leq i \leq m} \) that minimizes

\[
\sum_{i=1}^{m} \left( \hat{y}_i^{MQ} - t_i^{MQ} \right)^2
\]

and satisfies benchmarking and neutral shrinkage, i.e., is subject to the constraints:

1. \( \sum_{i=1}^{m} w_i t_i^{MQ} = c_1 \) (a reliable estimator of the overall population mean)
2. \( \sum_{i=1}^{m} w_i (t_i^{MQ} - t.)^2 = c_2 \) (a suitable measure of the variance between the areas)

\( t. = \sum_{i=1}^{m} w_i t_i^{MQ} \).
A Research Update (4)

- New linearization-based MSE estimator as alternative to use of parametric bootstrap and pseudo-linearization-based MSE estimation with robust estimators (Chambers, Chandra, Salvati and Tzavidis, 2014)

- A model-assisted approach and design consistent small area estimators based on the M-quantile small area model (Fabrizi, Salvati, Pratesi and Tzavidis, 2014)

\[
\hat{y}_{i}^{WMQ} = N_{i}^{-1} \sum_{j \in s_{i}} w_{ij} y_{ij} + \left( N_{i}^{-1} \sum_{j \in U_{i}} x_{ij}^{T} - N_{i}^{-1} \sum_{j \in s_{i}} w_{ij} x_{ij}^{T} \right) \hat{\beta}_{wq_{i}}
\]
A Research Update (5)

- Extension of M-quantile approach to GLM
  - Easy to compute alternative to a GLMM-based EB approach
  - Application to logistic M-quantile modelling and prediction very promising (Chambers, Salvati and Tzavidis, 2013)
  - M-quantile approach for counts is used for estimating the average number of visits to physicians for Health Districts in Central Italy (Tzavidis, Ranalli, Salvati, Dreassi and Chambers, 2014)
  - Semiparametric M-quantile regression for count data (Dreassi, Ranalli and Salvati, 2014)
  - Disease Mapping via Negative Binomial Regression M-quantiles (Chambers, Dreassi and Salvati, 2013)
Challenges for PRIN2014

Premise

- Linear, median, quantile, M-quantile regressions are semi-parametric models;

- Parametric distributional assumptions are useful for inferential purposes, especially in small samples (i.e. role of normality in linear regression);

- Asymmetric Laplace distribution is associated to quantile regression and used for testing goodness of fit (Koenker and Machado, 1999) and Bayesian inference (Yu and Moyeed, 2001; Sriram et al., 2013).
A new class of distributions

We introduced a new class of distributions (MAL):

$$f_q(u) = \frac{1}{B_q} \exp \left\{ - \rho_q(u) \right\}$$

with $u = \frac{y_i-x_i\beta_q}{s}$. The estimation of the M-quantile regression coefficient can then be represented as a MLE estimation problem.

Applications

- Model diagnostics (pseudo $R^2$, LRT test for linear hypotheses on $\beta_q$)
- ML estimation of the scale parameter $s$ and the tuning constant $c$
- Alternative approach to testing for the presence of (cluster) effects (i.e. inter cluster variation of best fitting $\beta_q$)
More applications

- Use of ‘MAL’ as likelihood for estimating M-quantile regression using parametric Bayesian techniques for complex models: MCMC, integrated nested Laplace approximations (INLA)

- Alternative variance estimation

- Easier extension to the estimation of complex parameters
Multinomial M-quantile regression models

- Multinomial modelling may be important for estimating labour market and other BES indicators.
- Extension of logistic regression methodology non-trivial as in multi-equation setting definition of observation or group specific quantile not straightforward.
- Multinomial M-quantile regression models. Dependent variable $y$ takes on $J+1$ integer values $(0, 1, ..., K)$ which correspond to different categories that are not overlapping and do not have a natural ordering:

\[ y = \begin{cases} 
0 & \text{with probability } Pr(Y = 0 | \mathbf{X}) \\
1 & \text{with probability } Pr(Y = 1 | \mathbf{X}) \\
& \vdots \\
K & \text{with probability } Pr(Y = K | \mathbf{X}) 
\end{cases} \]
Multinomial M-quantile regression models

Define the new binary variable $z_j$ for $k = 0, 1, \ldots, K$ as

$$y = \begin{cases} 
1 & Y = k \\
0 & \text{otherwise}
\end{cases}$$

Define

$$g_{q1}(X) = X\beta_{q1}$$

\ldots

$$g_{qk}(X) = X\beta_{qk}$$

\ldots

$$g_{qK}(X) = X\beta_{qK}$$

For $y = k$ estimating equations are therefore

$$\sum_{j=1}^{n} \left\{ w(x_j) v_{qk}^{1/2}(x_j) \left( \psi_q \left\{ \frac{Z_{jk} - \pi_{qk}(x_j)}{v_{qk}(x_j)} \right\} - E \left[ \psi_q \left\{ \frac{Z_{jk} - \pi_{qk}(x_j)}{v_{qk}(x_j)} \right\} \right] \right) x_j \right\}$$
Small area target parameters in the BES-3 and BES-4 subsets

BES-3: Labour Market

1: Tasso di occupazione 20-64 anni: Percentuale di occupati di 20-64 anni sulla popolazione totale di 20-64 anni.

2: Tasso di mancata partecipazione al lavoro: Percentuali di disoccupati di 15-74 anni + parte delle forze di lavoro potenziali di 15-74 anni (inattivi che non cercano lavoro nelle 4 settimane ma disponibili a lavorare) sul totale delle forze di lavoro 15-74 anni + parte delle forze di lavoro potenziali 15-74 anni (inattivi che non cercano lavoro nelle 4 settimane ma disponibili a lavorare).

Small area target parameters in the BES-3 and BES-4 subsets

BES-4: Economic well-being

- 2: Indice di disuguaglianza del reddito disponibile: Rapporto fra il reddito equivalente totale ricevuto dal 20% della popolazione con il più alto reddito e quello ricevuto dal 20% della popolazione con il più basso reddito.

- 10: Incidenza di persone che vivono in famiglie senza occupati: Percentuale di persone che vivono in famiglie dove è presente almeno un componente di 18-59 anni (con esclusione delle famiglie dove tutti i componenti sono studenti a tempo pieno con meno di 25 anni) dove nessun componente lavora o percepisce una pensione da lavoro sul totale delle persone che vivono in famiglie con almeno un componente di 18-59 anni.


References

- Pratesi, M., Ranalli, M.G. and Salvati, N. Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics*, 19, 687–701.