M-quantile models for Small Area Estimation

N. Salvati[1], E. Fabrizi[2]

[1]DEM, Pisa
[2]Piacenza

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

Introduction to Small Area Estimation (Cont´d)

- 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 $\mathbf{x}$
- Area level covariates $\mathbf{z}$
- Area random effect v
- Population / Sample / Non-sample $U/s/r$

Assumption 1

Linear relationship between y and $\mathbf{x}$, $\mathbf{z}$

Assumption 2

All small areas sampled

Assumption 3

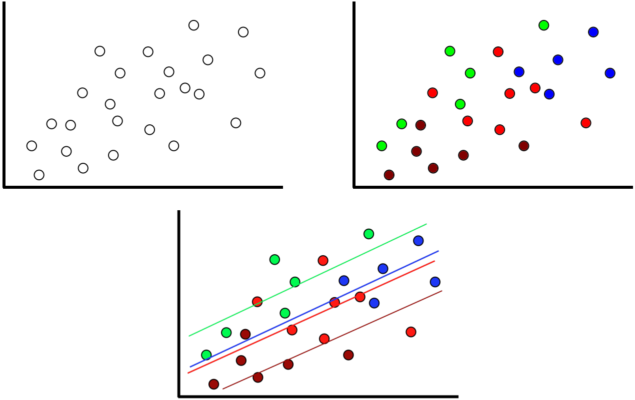
Small area means of $\mathbf{x}$ and $\mathbf{z}$ are known

The Industry Standard – EBLUP

Reference Rao (2003)
Working Model $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{e}$
Assumptions $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_v),\ \mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e)$
Empirical Best Linear Unbiased Predictor of $\bar{y}_i$ is

$$\hat{\bar{y}}_i = N_i^{-1}\Big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i}(\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{v}_i)\Big\} = N_i^{-1}\Big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i}\hat{y}_{ij}\Big\}$$

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 $f(y|\mathbf{X})$
- 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 $f(y|\mathbf{X})$
- A more complete picture is offered by modeling not only measures of central tendency of $f(y|\mathbf{X})$ 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 qth quantile of $f(y|\mathbf{X})$ is

$$Q_q(y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_\psi(q)$$

- Estimates of $\boldsymbol{\beta}_\psi(q)$'s are obtained via Iterative Weighted Least Squares (IWLS) by solving

$$\sum \psi_q \left\{ y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\psi(q) \right\} \mathbf{x}_i = 0$$

$$\hat{\boldsymbol{\beta}}_\psi(q) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

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

M-quantile models more formally (Cont´d)

- The influence function ψ_q is

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

- The influence function ψ is Huber Proposal II

$$\psi\{u\} = uI(-c < u < c) + csgn(u)$$

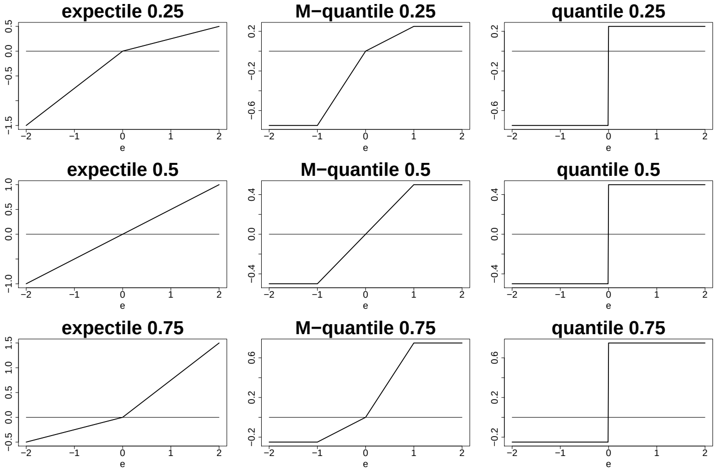
Why Regression M-quantiles?

- 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)



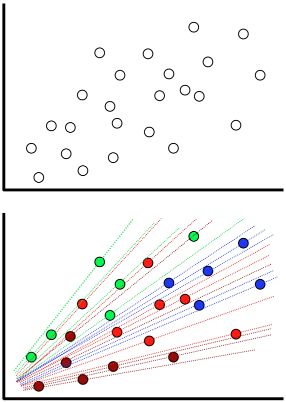
Using M-quantile Models in Small Area Estimation

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

$$\hat{\bar{y}}_i^{MQ} = N_i^{-1}\Big(\sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} \mathbf{x}_{ik}^T \hat{\beta}_{q_i}\Big) = N_i^{-1}\Big(\sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} \hat{y}_{ik}^{MQ}\Big)$$

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)

$$\hat{\bar{y}}_i^{NPMQ} = N_i^{-1}\Big(\sum_{j\in s_i} y_{ij} + \sum_{k\in r_i}\Big\{\mathbf{x}_{ik}^T\hat{\beta}_{q_i} + \mathbf{z}_{ik}^T\hat{\gamma}_{q_i}\Big\}\Big)$$

$$= N_i^{-1}\Big(\sum_{j\in s_i} y_{ij} + \sum_{k\in r_i}\hat{y}_{ik}^{NPMQ}\Big)$$

- Robust prediction of small area means and distributions (Tzavidis, Marchetti and Chambers, 2010)

A Research Update (2)

- Spatial version of the M-quantile models for small area estimation (Salvati, Tzavidis, Pratesi and Chambers, 2012)

$$\hat{\bar{y}}_i^{MQGWR} = N_i^{-1}\Big(\sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} \mathbf{x}_{ik}^T \hat{\beta}_{q_i}(g_{ik})\Big)$$

$$= N_i^{-1}\Big(\sum_{j \in s_i} y_{ij} + \sum_{k \in r_i} \hat{y}_{ik}^{MQGWR}\Big)$$

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 $\{\hat{\bar{y}}_i^{MQ}\}_{1\leqslant i\leqslant m}$ of the small area means, we look for a new set of estimators $\{t_i^{MQ}\}_{1\leqslant i\leqslant m}$ that minimizes

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

  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)

$$\hat{\bar{y}}_i^{WMQ} = N_i^{-1} \sum_{j \in s_i} w_{ij} y_{ij} + \left(N_i^{-1} \sum_{j \in U_i} \mathbf{x}_{ij}^T - N_i^{-1} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij}^T \right) \hat{\boldsymbol{\beta}}_{w q_i}$$

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 $u = \frac{y_i - \mathbf{x}_i \boldsymbol{\beta}_q}{s}$. 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 $\boldsymbol{\beta}_q$)
- 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 $\boldsymbol{\beta}_q$)

More applications

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

- Alternative variance estimation
- Easier extension to the estimation of complex parameters

Multinomial M-quantile regression models

- Multinomila 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:

$$y = \begin{cases} 0 & \text{with probability } Pr(Y = 0|\mathbf{X}) \\ 1 & \text{with probability } Pr(Y = 1|\mathbf{X}) \\ \vdots & \qquad\qquad\vdots \\ K & \text{with probability } Pr(Y = K|\mathbf{X}) \end{cases}$$

Multinomial M-quantile regression models

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

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

Define
$g_{q1}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_{q1}$
$\ldots$
$g_{qk}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_{qk}$
$\ldots$
$g_{qK}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_{qK}$
For $y = k$ estimating equations are therefore

$$\sum_{j=1}^{n} \left\{ w(\mathbf{x}_j) v_{qk}^{1/2}(\mathbf{x}_j) \left(\psi_q \left\{ \frac{z_{jk} - \pi_{qk}(\mathbf{x}_j)}{v_{qk}(\mathbf{x}_j)} \right\} - E\left[\psi_q \left\{ \frac{z_{jk} - \pi_{qk}(\mathbf{x}_j)}{v_{qk}(\mathbf{x}_j)} \right\} \right] \right) \mathbf{x}_j \right\}$$

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).
- 9: Rapporto tra tasso di occupazione delle donne di 25-49 anni con figli in età prescolare e delle donne senza figli: Tasso di occupazione delle donne di 25-49 anni con almeno un figlio in età 0-5 anni sul tasso di occupazione delle donne di 25-49 anni senza figli per 100.

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

- Chambers, R. and Dunstan, R. (1986). Estimating distribution function from survey data, *Biometrika*, **73**, 597–604.
- Chambers, R. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. *Biometrika*, **93**, 255–268.
- Chambers, R., Chandra, H. and Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **37**, 153–170.
- Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, **81**, 1063–1069.
- Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, **76**, 47–69.
- Chambers, R.L. and Clark, R.G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press: Oxford.
- Chambers, Dreassi, E. and Salvati, N. (2013). Disease Mapping via Negative Binomial Regression M-quantiles. arXiv:1310.3403v1 [stat.ME].
- Dreassi, E., N., Ranalli, M.G. and Salvati, N. (2014). Robust small area prediction for counts. *Statistical Methods in Medical Research*. To appear.
- Fabrizi, E., Salvati, N., Pratesi, M. and Tzavidis, N. Outlier robust model-assisted small area estimation. *Biometrical Journal*, **56**, 157–175.
- Fabrizi, E., Salvati, N. and Pratesi, M. Constrained Small Area Estimators Based on M-quantile Methods. *Journal of Official Statistics*, **28**, 89–106.
- Koenker, R. and Machado, J.A.F. (1999). Goodness of fit and related inference processes for quantile regression, *Journal of the American Statistical Association*, **94**, 1296–1310.

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.
- Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. Small area estimation via M-quantile geographically weighted regression. *TEST*, **21**, 1–28.
- Sinha, S.K. and Rao, J.N.K. (2009) Robust small area estimation. *Canadian Journal of Statistics*, **37**, 381–399.
- Sriram K., Ramamoorty R.V. and Ghosh P. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified Asymmetric Laplace density, *Bayesian Analysis*, **8**, 1–26.
- Tzavidis, N., Chambers, R., Salvati, N. and Chandra, H. (2012). Small area estimation in practice: An application to agricultural business survey data. *Journal of the Indian Society of Agricultural Statistics*, **66**, 213–228.
- Tzavidis, N., Ranalli, M.G., Salvati, N., Dreassi, E. and Chambers, R. (2014). Robust small area prediction for counts. *Statistical Methods in Medical Research*, doi:10.1177/0962280214520731.
- Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, **52**, 167–186
- Yu K. and Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics and Probability letters*, **54**, 437–447.