Three-level M-quantile model for poverty estimation

Stefano Marchetti[1,2], Nicola Salvati[1,2]

1. Department of Economics and Management, University of Pisa
2. Interuniversity Tuscany Centre for Advanced Statistics for Equitable and Sustainable Development "Camilo Dagum"

SIS 2018
Palermo, 20-22 June 2018

Structure of the Presentation

Part I

Introduction

SAE poverty mapping

- Small area estimation (SAE) aims to allow efficient estimation of population characteristics of domains (typically associated with an administrative geography of interest) for which direct estimation is unreliable (Rao, 2003)
- Standard SAE methods deal with totals or means of target variables (linear functions)
- Area-specific poverty indicators are complex non-linear functions of small area income distributions (Betti et al., 2006) and require modified SAE methods
- We propose a three-level M-quantile model to estimate non-linear small area characteristics

FGT poverty measure

Poverty indicator for area d

$$FGT_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{t - W_{dj}}{t} \right)^{\alpha} I(W_{dj} \leq t)$$

- t is the poverty line (usually $0.6 \times \text{median}(W)$)
- N_d is the number of households in area d
- W_{dj} welfare value (usually income) of household j in area d
- $\alpha = 0$ FGT is the HCR (P_d), $\alpha = 1$ FGT is the PG

Part II

Three-level M-quantile linear model

M-quantile (Mq) model, short review

- The M-quantile Q_q $(q \in (0,1))$ of a random variable Y is the solution of the equation $\int \psi_q(y-\theta_q)dF(Y) = 0$
- $\psi_q(u) = 2\psi(u)[(1-q)I(u<0) + qI(u \geq 0)]$
- $\psi(u)$ is an opportunely chosen influence function
- The Mq Q_q of the conditional distribution $Y|\boldsymbol{X}$ is the solution of the equation $\int \psi_q(y - Q_q(\boldsymbol{x}))dF(Y|\boldsymbol{X}) = 0$
- Regression model (linear): $Q_q(x) = \boldsymbol{X}^T\boldsymbol{\beta}_\psi(q)$
- $\forall\, y_{dj}$ $(j \in s_d, d = 1, \dots, D)$ $\exists\, q_{dj} \in [0,1] \,|\, y_{dj} = Q_{q_{dj}}(x) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta}_\psi(q)$
- $N_d^{-1}\sum_{j=1}^{N_d} q_{dj} = \theta_d$ (Area d M-quantile)
- Mq small area model: $y_{dj} = \boldsymbol{x}_{dj}^T\boldsymbol{\beta}_\psi(\theta_d) + \epsilon_{dj}$
- Mq small area model can mimic the two-level mixed effect model

Three-level M-quantile model

- Extend the Mq model to a three-level Mq model to take into account variability for groups *within* areas
- D areas and C clusters (clusters are partitions of areas)
- C_d clusters in area d, N_{c_d} units in each cluster
- A three-level Mq model can be defined as follows

$$y_{dcj} = \boldsymbol{x}_{dcj}^T \boldsymbol{\beta}_\psi(\theta_d) + \boldsymbol{z}_{dcj}^T \boldsymbol{\gamma}_\psi(\phi_{dc}) + \epsilon_{dcj},$$

- y_{dcj} continuous variable (unit j, cluster c, area d)
- $\boldsymbol{x}_{dcj}$, $\boldsymbol{z}_{dcj}$ p- and q-vector of auxiliary variables (known for N units)
- θ_d is the area d Mq coefficient, ϕ_{dc} is the cluster $c \in d$ Mq coefficient
- $\boldsymbol{\beta}_\psi$ and $\boldsymbol{\gamma}_\psi$ are the vectors of area-level and cluster-level regression coefficients
- ϵ_{dcj} is a unit error term (no distributional assumptions)

Three-level M-quantile model: parameters estimation

1. Starting from the Mq linear model

$$y_{dcj} = \boldsymbol{x}_{dcj}^T \boldsymbol{\beta}_\psi(\theta_d) + u_{dcj},$$

estimate θ_d and $\boldsymbol{\beta}_\psi(\theta_d)$ according to the Mq approach to SAE

2. Compute the residuals $\hat{u}_{dcj} = y_{dcj} - \boldsymbol{x}_{dcj}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d)$
3. Using residuals $\hat{u}_{dcj}$ as the target variable in the Mq linear model

$$\hat{u}_{dcj} = \boldsymbol{z}_{dcj}^T \boldsymbol{\gamma}_\psi(\phi_{dc}) + \epsilon_{dcj},$$

estimate ϕ_{dc} and $\boldsymbol{\gamma}_\psi$ by using again the Mq approach to SAE ($\hat{u}_{dcj}$ target, ϕ_{dc} cluster Mq coefficient, $\boldsymbol{\gamma}_\psi(\phi_{dc})$ regression coefficients); residual of Mq three-level is $\hat{\epsilon}_{dcj} = y_{dcj} - \boldsymbol{x}_{dcj}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) - \boldsymbol{z}_{dcj}^T \hat{\boldsymbol{\gamma}}_\psi(\hat{\phi}_{dc})$

If there are no auxiliary variables at cluster level, it is possible to use the same set of auxiliary variables in step 1 and 3, $\boldsymbol{z}_{dcj} = \boldsymbol{x}_{dcj}$ for all units (Mq three-level model mimics a three-level linear mixed model)

Three-level M-quantile model: motivation

- It is possible to mimic a three-level mixed model by using a common set of auxiliary variables
- Unit level M-quantile coefficients are computed for each unit in the sample, then ‘aggregated’ (averaged) at area level to obtain an Mq ‘area’ coefficient
- Then, model parameters are estimated and residuals are computed
- The Mq model in its part $\boldsymbol{x}_{dcj}^T \hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_d)$ captures the variability that is explained by the auxiliary variable together with the variability that is explained by the hierarchical structure in the data (e.g. area level hierarchy)

Three-level M-quantile model: motivation

- The residuals of this model, $\hat{u}_{dcj}$, include residual variability due to unobserved variables, unobservable factors and other sources of variability related to different hierarchies in the data
- This last source of variability can be captured in step 3 of the proposed procedure
- Using the same set of auxiliary variables $\boldsymbol{x}_{dcj}^T$, on the residuals $\hat{u}_{dcj}$, obtained in step 2, it is possible to capture residual variability explained by auxiliary variables together with between-cluster variability (cluster level hierarchy) by the 'aggregation' mechanism in the M-quantile small area model, i.e. $\boldsymbol{x}_{dcj}^T \hat{\gamma}_\psi(\hat{\phi}_{dc})$.
- Model-based simulations show a very high correlation between area random effects of three-level mixed model and area pseudo-random effects of the Mq three-level; the same for cluster random errors

Part III

SAE using three-level Mq model

SAE under three-level Mq model: MC approach

- Our idea is to use a Monte Carlo approach to micro-simulate the population values of the target variable by means of three-level Mq model
- By this approach is then possible to obtain estimates of the target parameters of interest in the desired areas or domains
- In this work the goal is to estimate a parameter that is function of the target variable y (continuous) $\rightarrow h(y)$
- The data required to apply the proposed method are
  - a random sample drawn from the target population (observed y_j, $\boldsymbol{x}_j$, $j \in s$)
  - the auxiliary variables for all the units of the population ($\boldsymbol{x}_j$, $j \in U$)
  - area and cluster indicators for sample and population units

SAE under three-level Mq model

Working environment

- D sampled areas out of D areas and m sampled clusters out of M clusters, $M - m$ out of sample clusters
- Areas are partitions of the population; clusters are partitions of the areas
- In m clusters of size N_{d_c} a random sample of n_{d_c} units is available
- s_d and r_d are set of sampled and non-sampled units in area d
- Parameter of interest: $h_d(y) = N_d^{-1}\{\sum_{j \in s_d} h(y_{dcj}) + \sum_{j \in r_d} h(y_{dcj})\}$
- $\sum_{j \in r_d} h(y_{dcj})$ is unknown and have to be estimated
- The predictor takes the form $\hat{h}_d(y) = N_d^{-1}\{\sum_{j \in s_d} h(y_{dcj}) + \sum_{j \in r_d} \hat{h}(\hat{y}_{dcj})\}$
- $\hat{h}(\hat{y}_{dcj})$ is the predictor of $h(y_{dcj})$ given by $E[h(y_{dcj})|\boldsymbol{y}_s]$

SAE under three-level Mq model: MC approach

To obtain $\hat{h}_d(y)$ we propose a Monte Carlo approximation as follows

1. Estimate model $y_{dcj} = \boldsymbol{x}_{dcj}^T \boldsymbol{\beta}_\psi(\theta_d) + \boldsymbol{z}_{dcj}^T \boldsymbol{\gamma}_\psi(\phi_{dc}) + \epsilon_{dcj}$ by using sample data
2. Generate a synthetic population, predicting non-sampled units

   $$\hat{y}_{dcj}^{syn} = \boldsymbol{x}_{dcj}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) + \boldsymbol{z}_{dcj}^T \hat{\boldsymbol{\gamma}}_\psi(\hat{\phi}_{dc}) \quad c = 1, \ldots, m_d, j = 1, \ldots, N_{d_c} - n_{d_c},$$

   for out of sample clusters $\hat{y}_{dc_{out}j}^{syn} = \boldsymbol{x}_{dcj}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) + \boldsymbol{z}_{dcj}^T \hat{\boldsymbol{\gamma}}_\psi(0.5)$
3. Generate k MC values for non-sampled units

   $$\hat{y}_{jcd}^k = \hat{y}_{dcj}^{syn} + u_d^* + v_{cd}^* + \epsilon_{dcj}^*,$$

   - ϵ_{dcj}^* are sampled with replication from model residuals $\hat{\epsilon}_{dcj}$
   - u_d^* are sampled with replication from pseudo-area effects $\boldsymbol{x}_{dcj}^T \{\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) - \hat{\boldsymbol{\beta}}_\psi(0.5)\}$
   - v_{dc}^* are sampled with replication from pseudo-cluster effects $\boldsymbol{z}_{dcj}^T \{\hat{\boldsymbol{\gamma}}_\psi(\hat{\phi}_{dc}) - \hat{\boldsymbol{\gamma}}_\psi(0.5)\}$

SAE under three-level Mq model: MC approach

4 compute the target parameter(s) on the kth MC population

$$\hat{h}_d^k(y) = N_d^{-1}\left\{\sum_{j \in s_d} h(y_{dcj}) + \sum_{j \in r_d} h(\hat{y}_{dcj}^k)\right\}$$

5 repeat steps 3 and 4 K times and then estimate $h_d(y)$ by averaging over the K MC populations

$$\hat{h}_d(y) = K^{-1} \sum_{k=1}^{K} \hat{h}_d^k(y)$$

SAE under three-level Mq model: MC approach

- The disturbances that are added in step 3 are justified since y_{dcj}^{syn} is the expected value under the three-level model of unknown quantity y_{dcj} $(j \in r_d)$ that has an unknown distribution
- By adding a pseudo-area and a pseudo-cluster error as well as a unit level error we mimic non-parametrically the unknown distribution of y_{dcj} $(j \in r_d)$
- It is an approach similar in spirit to that used by Molina and Rao (2010) and Marhuenda et al. (2017).
- MSE estimation can be obtained using the bootstrap technique proposed by Marchetti et al. (2018)

Model-based simulation

- The model-based simulation experiment is designed to compare Mq estimators derived from the proposed three-level Mq model with Mq estimators derived from the traditional Mq model, which accounts for only one hierarchical structure in the data
- We also compare Mq estimators with empirical best predictors (EBP) derived from the three-level mixed model and EBPs derived from the traditional two-level mixed model
- Populations are generated according to Marhuenda et al. (2017), but focusing only on domain estimation
- In this setting we expect EBPs to perform the best.
- The use of this setting is useful to check the performance of Mq estimators with respect to each other and compared with EB estimators, which are characterised by the best performance when random effects are normally distributed.

Data generation process

- $N = 20000$ units
- $D = 40$ areas, $N_d = 500, d = 1, \ldots, D$
- Each area is divided into $M_d = 10$ clusters, $N_{cd} = 50, c = 1, \ldots, M_d$
- Auxiliary variables:
  - $x_{1,dcj} \sim Bin(1, P_{1,dc})$ with $P_{1,dc} = 0.2 + 0.4d/D + 0.4c/M_d$
  - $x_{2,dcj} \sim Bin(1, P_{2,dc})$ with $P_{2,dc} = 0.2, c = 1, \ldots, M_d, d = 1, \ldots, D$
- Area effects $u_d \sim N(0, \sigma_u^2)$ (changes according to different scenarios)
- Cluster effects $v_{dc} \sim N(0, \sigma_v^2)$ (changes according to different scenarios)
- Unit-level errors $e_{dcj} \sim N(0, \sigma_e^2 = 0.25)$

Data generation process

- Target variable W_{dcj}, but we model $Y_{dcj} = \log W_{dcj}$
- The population of Y values is generated from the following model

$$Y_{dcj} = 3 + 0.03x_{1,dcj} - 0.04x_{2,dcj} + u_d + v_{dc} + e_{dcj}$$
$$j = 1, \dots, N_{d_c}, c = 1, \dots, M_d, d = 1, \dots, D$$

- Poverty line is set as $z = 0.6 \times median(W_{dcj})$, with $W_{dcj} = \exp(Y_{dcj})$
- From the generated populations we draw simple random samples without replacement, s_{d_c}
- From each of the $m_d = M_d = 10$ clusters within area d, with size $n_{d_c} = 5 < N_{d_c} = 50$
- Area sample size is $n_d = 50 < N_d = 500, d = 1, \dots, D$

Data generation process: scenarios

- To represent the various scenarios that can be found in real applications we generate the population assuming different values of random area and cluster effects
- $\sigma_v = 0.1$, $\sigma_u \in \{0, 0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$
  - $\rho_{area} \in \{0, 0.002, 0.009, 0.037, 0.080, 0.133, 0.194, 0.257\}$
- $\sigma_u = 0.2$, $\sigma_v \in \{0, 0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$
  - $\rho_{area} \in \{0.138, 0.138, 0.137, 0.133, 0.128, 0.121, 0.113, 0.105\}$

Simulation

- We run $K = 1000$ MC replications
- Population $\Omega_k = Y^k_{dcj}, x^k_{1,dcj}, x^k_{2,dcj}$, $j = 1, \ldots, N_{d_c}$, $c = 1, \ldots, M_d$, $d = 1, \ldots, D$
- True values of the area Head Count Ratio (HCR), $P^k_d = N_d^{-1} \sum_{j \in \Omega^k_d} \{z^k - \exp(Y^k_{dj})\}(z^k)^{-1} I(\exp(Y^k_{dj}) \leq z^k)$; $z^k = 0.6 \times median(\exp\{Y^k_{dj}\})$
- Sample from Monte-Carlo populations, $s^k = \{s^k_1, \ldots, s^k_d, \ldots, s^k_D\}$
- Models: three- and two-level Mq models and the three- and two-level mixed models
- Estimators ($\hat{P}^k_d$): MQ3, MQ2, EB3, EB2, Dir

Evaluation

We computed values of empirical bias and empirical MSE of these estimators for areas as follows

$$B(\hat{P}_d^k) = K^{-1} \sum_{k=1}^{K} (\hat{P}_d^k - P_d^k)$$

$$MSE(\hat{P}_d^k) = K^{-1} \sum_{k=1}^{K} (\hat{P}_d^k - P_d^k)^2$$

Main Results

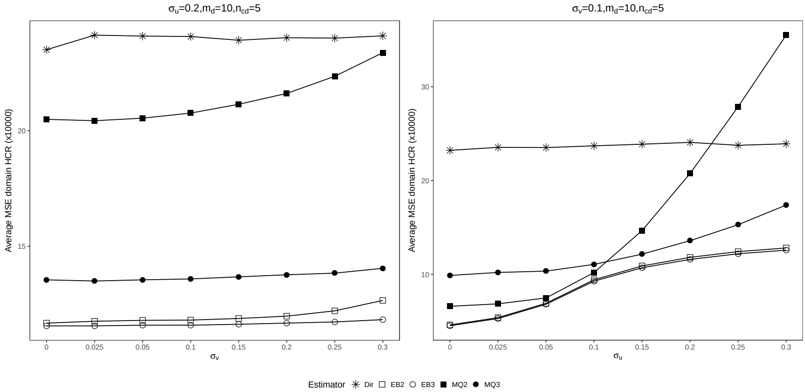

Figure: Average $MSE_d \times 10^4$ across areas for the estimators of area-level HCRs, with $\sigma_u = 0.2$ (left) and $\sigma_v = 0.1$ (right).

Main Results

Table: Average $B_d \times 10^2$ across areas for the estimators of area-level HCRs, for $\sigma_u = 0.2$ with varying σ_v and $\sigma_v = 0.1$ with varying σ_u.

	$\sigma_u = 0.2$							
$\sigma_v =$	0	0.025	0.05	0.1	0.15	0.2	0.25	0.3
MQ2	−0.23	−0.29	−0.29	−0.29	−0.27	−0.26	−0.27	−0.28
MQ3	−0.10	−0.11	−0.11	−0.10	−0.05	0.02	0.11	0.24
EB2	0.04	−0.01	−0.01	−0.01	0.00	0.00	0.00	−0.01
EB3	0.04	−0.02	−0.04	−0.02	−0.01	0.00	−0.01	−0.02
Dir	0.02	−0.06	−0.07	−0.05	−0.05	−0.01	0.01	0.02
	$\sigma_v = 0.1$							
$\sigma_u =$	0	0.025	0.05	0.1	0.15	0.2	0.25	0.3
MQ2	−0.08	−0.14	−0.15	−0.19	−0.24	−0.29	−0.29	−0.28
MQ3	−0.21	−0.24	−0.22	−0.26	−0.25	−0.10	0.11	0.23
EB2	−0.01	−0.02	−0.02	−0.02	−0.01	−0.01	0.01	0.00
EB3	0.00	−0.02	−0.02	−0.02	−0.02	−0.02	0.00	0.00
Dir	−0.03	−0.05	−0.07	−0.04	−0.04	−0.05	−0.04	−0.01

Discussion

Settings with fixed $\sigma_u = 0.2$

- EB3, EB2 and MQ3 perform better than MQ2 and Dir
- Best results EB3, close to EB2 (expected)
- $\sigma_v \uparrow$ $MSE \downarrow$ for all the estimators, but Dir
- ρ_{area} declines from about 14% to 11%, $\rho_{cluster}$ increases from about 14% to 34% (0.138, 0.14, 0.145, 0.167, 0.2, 0.24, 0.291, 0.342)
- $\rho_{cluster} \uparrow$ MQ2 performs poorly in terms of variability
- The same happens in the case of the EB estimators, EB3 outperforms EB2 as $\rho_{cluster} \uparrow$ (difference of MSE between EB2 and EB3 is minimal)
- Better performance of MQ3 compared to MQ2 is confirmed for bias

Discussion

Settings with fixed $\sigma_v = 0.1$

- ρ_{area} increases from 0% to about 26%, $\rho_{cluster}$ increase from about 4% to 29%, (0.038,0.041,0.048,0.074,0.115,0.167,0.225,0.286)
- For values of $\sigma_u < 0.1 \implies \rho_{area} < 3.7\%$ and $\rho_{cluster} < 7\%$ MQ2 performs better than MQ3, EB2 has the best performance
- When $\sigma_v = \sigma_u = 0.1$ the average MSE of EB2, MQ2 and MQ3 are very close to each other
- When ρ_{area} and $\rho_{cluster}$ increase from the above values, then MQ3 performs better and outperforms MQ2 in terms of MSE
- MQ2 is a little bit better than MQ3 in term of bias in all these settings, but $\sigma_u \geq 0.2$
- The EB2 and EB3 estimators perform best, as expected in this simulation framework

Discussion

- The model-based simulation experiments reveal an overall good performance of MQ3 with respect to MQ2
- When cluster and area intraclass correlations are small, like $\rho_{cluster} < 10\%$ and $\rho_{area} < 4\%$, MQ2 can outperform MQ3 in terms of variability
- When cluster and area intraclass correlations are about 10% or bigger, then MQ3 is better than MQ2, which performs poorly for high values of $\rho_{cluster}$.

Discussion

- More details are in Marchetti, S., Beresewicz, M., Salvati, N., Szymkowiak, M. and Wawrowski, L. (2018). The use of a three-level M quantile model to map poverty at local administrative unit 1 in Poland, *J. R. Statist. Soc. A*
- Design-based simulation based on Poland data shows a better performance of MQ3 and MQ2 than EB3 and EB2, and MQ3 performs the best
- Design-based simulation shows the validity of the bootstrap techinique proposed in the paper
- Application to Poland data shows the best results in term of CV for MQ3 on poverty incidence and average income

Part IV

Concluding remarks

Conclusions

- We proposed a three-level Mq model
- We proposed a MC technique based on the three-level Mq model to estimate poverty incidence (easily extensible to a family of population parameters)
- We compared MQ3, MQ2, EB3, EB2 and Dir by MC model-based simulations under different intraclass correlation scenarios
- MQ3 shows a better efficiency with respect to MQ2 when area and cluster intraclass correlation is greater than about 10%
- Design-based simulations and application based on Poland data show the validity of MQ3