The Small Area Estimation of the quantiles of income distribution

S. Marchetti C. Giusti N. Salvati M. Pratesi

Department of Statistics and Mathematics Applied to Economics, University of Pisa

Outline

Part I

Industry Standard for SAE and Alternative Estimators (M-quantile Approach)

The Industry Standard for Small Area Estimation: Mixed Effects Models that Include Random Area Effects

Concept

Include random area-specific effects to account for the between area variation beyond that explained by the variation in model covariates

Notation: (j =area, i =individual)

- Variable of interest: y_{ij}
- Focus on unit level covariate information: $\mathbf{x}_{ij}$
- Area level random effect: γ_j
- Random error: ϵ_{ij}

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\gamma_j + \epsilon_{ij},\ i = 1, ..., n_j,\ j = 1, ...d$$

Estimator of Small Area Mean

$$\hat{m}_j = N_j^{-1}\left(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \mathbf{z}'_{ij}\hat{\gamma}_j\right)$$

An Alternative Approach: Using M-quantile Models to Measure Area Effects

- Individual level data on y and $\mathbf{x}$
- Each sample pair (x, y) lies on one and only one M-quantile line
- The q-value of this line = **M-quantile coefficient** or q value of the corresponding sample unit
- Calculate an *M-quantile coefficient* for each area j by suitably **averaging** the q values of each sampled individual in that area. Denote this average **area-specific q-value** by θ_j
- Estimate the area specific target parameter by fitting an M-quantile model for each area at $\hat{\theta}_j$

$$y_{ij} = Q_{\theta_j}(y|\mathbf{x}) = \mathbf{x}'_{ij}\boldsymbol{\beta}_{\psi}(\theta_j)$$

Part II

Model-Based Estimators of the Distribution Function

Model-Based Estimators of the Finite Population Distribution Function

- The empirical distribution function for a finite population is

$$F(t) = \frac{1}{N}\left(\sum_{i \in s} I(y_i \leq t) + \sum_{i \in r} I(y_i \leq t)\right)$$

- $\Omega = \{1, \ldots, N\}$ population units
- $s = \{1, \ldots, n\}$ sampled units
- $r = \{\Omega - s\}$ non sampled units

Model-Based Distribution Function Estimators for Finite Population

Distribution Function Estimator: Naïve Approach

$$\hat{F}(t) = \frac{1}{N}\left(\sum_{i\in s} I(y_i \leq t) + \sum_{k\in r} I(\hat{y}_k(\mathbf{x}_k) \leq t)\right)$$

Distribution Function Estimator: Chambers-Dunstan Approach

$$\hat{F}(t)_{CD} = \frac{1}{N}\left[\sum_{i\in s} I(y_i \leq t) + \frac{1}{n}\sum_{k\in r}\sum_{i\in s} I\left(\hat{y}_k(\mathbf{x}_k) + (y_i - \hat{y}_i) \leq t\right)\right]$$

Where $y_i = y_i(\mathbf{x}_i) = f(\mathbf{x}_i) + \varepsilon_i$, $i \in \Omega$, $\mathbf{x}_i$ auxiliary variables vector for the unit i that is known for all the population units

Model-Based Small Area Distribution Function Estimators for Finite Population

An estimator of the Chambers-Dunstan small area distribution function can be defined under different SAE models

- Mixed models (unsuitable for the Chambers-Dunstan approach)

$$\hat{F}(t)_{j,CD} = \frac{1}{N_j}\left[\sum_{i\in s_j} I(y_i \leq t) + \frac{1}{n_j}\sum_{k\in r_j}\sum_{i\in s_j} I\left(\mathbf{x}_k'\hat{\beta} + \mathbf{z}_k'\hat{\gamma}_j + (y_i - \hat{y}_i) \leq t\right)\right]$$

- M-Quantile models

$$\hat{F}(t)_{j,CD} = \frac{1}{N_j}\left[\sum_{i\in s_j} I(y_i \leq t) + \frac{1}{n_j}\sum_{k\in r_j}\sum_{i\in s_j} I\left(\mathbf{x}_k'\hat{\beta}_\psi(\hat{\theta}_j) + (y_i - \hat{y}_i) \leq t\right)\right]$$

An Estimator of the Small Area Quantile

- The pth quantile for small area j ($\hat{q}_j(p)$) can be obtained by numerically solving the integral

$$\hat{q}_j(p) : \int_{-\infty}^{\hat{q}_j(p)} d\hat{F}_{j,CD}(t) = p$$

Part III

A Mean Squared Error Estimator of the Chambers-Dunstan Small Area Distribution Function

A Mean Squared Error Estimator of the Chambers-Dunstan Small Area Distribution Function

- Our MSE estimator for the small area distribution function is based on the bootstrap method proposed by Lombardia et all. (2003).
- In this work we extended the Lombardia et all. (2003) bootstrap method to the small area estimation problem under the M-quantile approach

A Mean Squared Error Estimator of the Chambers-Dunstan Small Area Distribution Function

- Let $b = (1, \ldots, B)$, where B is the number of bootstrap populations
- Let $r = (1, \ldots, R)$, where R is the number of bootstrap samples
- Let $\Omega = (y_k, \mathbf{x}_k)$, $k \in (1, \ldots, N)$, be the target population
- By $\cdot^*$ we denote bootstrap quantities
- $\hat{F}(t)_{j,CD} = \hat{F}(t)$ denotes the Chambers-Dunstan estimator of the distribution function of the small area j
- Let y be the study variable that is known only for sampled units and let $\mathbf{x}$ be the vector of auxiliary variables that is known for all the population units
- Let $s = (1, \ldots, n)$ be a within area simple random sample of the finite population $\Omega = \{1, \ldots, N\}$
- Let $\hat{y}_{ij} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_{\psi_p}(\hat{\theta}_j)$ be the estimate of the M-quantile regression model based on sample s
- Let $y_{ij} - \hat{y}_{ij} = e_{ij}$ be the residual of the unit i in area j

A Mean Squared Error Estimator of the Chambers-Dunstan Small Area Distribution Function

- Generate B bootstrap populations of dimension N, Ω^{*b}
  - i $y^*_{kj} = \mathbf{x}'_{kj}\hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_j) + e^*_{kj}$, $k = (1, \ldots, N)$
  - ii e^*_{kj} are obtained by sampling with replacement residuals e_{ij}
  - iii residuals can be sampled from the empirical distribution function or from a smoothed distribution function
  - iv we can consider all the residuals $(e_i, i = 1, \ldots, n)$, that is the unconditional approach or only area residuals $(e_{ij}, i = 1, \ldots, n_j)$, that is the conditional approach.
- From every bootstrap population we draw R samples of size n without replacement

A Mean Squared Error Estimator of the Chambers-Dunstan Small Area Distribution Function

- From the B bootstrap populations and from the R samples drawn from every bootstrap population we can estimate the mean squared error of the Chambers-Dunstan estimator of the distribution function

Bias

$$\hat{E}\left[\hat{F}^*(t) - F^*(t)\right] = B^{-1}\sum_{b=1}^{B} R^{-1}\sum_{r=1}^{R}\left(\hat{F}^{*br}(t) - F^{*b}(t)\right)$$

Variance

$$\widehat{Var}\left[\hat{F}^*(t) - F^*(t)\right] = B^{-1}\sum_{b=1}^{B} R^{-1}\sum_{r=1}^{R}\left(\hat{F}^{*br}(t) - \bar{\hat{F}}^{*br}(t)\right)^2$$

where

- $F^{*b}(t)$ is the distribution function of the bth bootstrap population
- $\hat{F}^{*br}(t)$ is the Chambers-Dunstan estimate for $F^{*b}(t)$ estimated using the rth sample drown from the bth bootstrap population
- $\bar{\hat{F}}^{*br}(t) = R^{-1}\sum_{r=1}^{R}\hat{F}^{*br}$

Part IV

Simulation Results

Simulation Design

- Population data has been generated from a random intercept model

$$y_{ij} = 1 + x_{ij} + \gamma_j + \epsilon_{ij}, \quad j = 1, \ldots, 30, \quad i = 1, \ldots, N_j$$

- $70 \leq N_j \leq 440$ and $N = 7430$, $7 \leq n_j \leq 44$ and $n = 743$
- $x_{ij} \sim N(\mu_j, \sigma_x^2 = 1)$, $22 \leq \mu_j \leq 93$
- $\gamma_i \sim N(0, \sigma_u^2 = 1), \epsilon_{ij} \sim N(0, \sigma_e^2 = 4)$
- $\gamma_i \sim \chi^2(3), \epsilon_{ij} \sim \chi^2(3)$
- $\gamma_i \sim t(3), \epsilon_{ij} \sim t(3)$
- The target parameters are the 25th, 50th, 75th small area percentiles

Model Based Simulation

Table: Relative Bias (RB), Absolute Relative Bias (ARB) and Relative Root Mean Squared Error (RRMSE) of the RMSE estimator of the 50th percentile, summarized over areas and simulations. Empirical Unconditional Approach.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Normal						
RB(%)	−6.24	−2.12	−0.23	−0.68	0.59	5.86
ARB(%)	6.20	7.12	7.89	8.21	9.03	11.47
RRMSE(%)	5.95	7.28	10.11	11.36	14.16	25.09
χ^2						
RB(%)	−5.22	−1.51	0.59	0.50	2.79	5.44
ARB(%)	6.34	7.55	8.41	8.33	9.03	10.81
RRMSE(%)	5.55	7.08	9.39	10.78	13.59	22.43
t						
RB(%)	−4.97	−1.63	0.23	0.46	2.63	6.14
ARB(%)	6.75	7.57	8.97	9.26	10.98	16.43
RRMSE(%)	4.20	5.35	8.21	9.69	13.08	35.24

Part V

An Application: Estimating the Income Quartiles for Lombardia, Toscana and Campania Provinces

Estimating the Income Quartiles for Campania, Toscana and Lombardia Provinces

- Data on the equivalised income in 2007 are available from the EU-SILC survey 2008 for 1489 households in the 10 Tuscany Provinces, for 1317 households in the 5 Campania Provinces and for 2214 households in the 11 Lombardia Provinces
- A set of explanatory variables is available for each unit in the population from the Population Census 2001
- We employ M-quantile models to estimate the first, the second (median), and the third quartile of the household equivalised income.

Model Specifications

- The selection of covariates to fit the small area models relies on prior studies of poverty assessment
- The following covariates have been selected:
  - household size (integer value)
  - ownership of dwelling (owner/tenant)
  - age of the head of the household (integer value)
  - years of education of the head of the household (integer value)
  - working position of the head of the household (employed / unemployed in the previous week)
  - gender of the head of the household