Estimating Inequality and Poverty Indexes at a Local Level

Stefano Marchetti and Caterina Giusti

Department of Economics and Management, University of Pisa

Kickoff Meeting PRIN
Roma, 12-13 January 2014

Structure of the Presentation

1. Motivation
2. M-quantile Models
3. M-quantile Estimator of Poverty Indexes
4. M-quantile Estimator of the Gini Coefficient
5. M-quantile Estimator of the Theil Index
6. MSE estimation of small area Poverty and Inequality Indexes
7. Monte Carlo Method for Poverty and Inequality Indexes
8. Concluding remarks

Part I

Motivation

Motivation

Problem

Estimate indicators for social exclusion at the small area level (areas for which direct estimates are not reliable)

How to handle the problem

BES n.4 “Benessere Economico” defines a set of indicators to monitor social exclusion

Aim

Estimates BES n. 4 indicators at local level (small area level)

Motivation

Required methodology to estimate selected indicators under M-quantile approach

- for small area means and totals $\rightarrow$ present in literature
- for small area income distribution $\rightarrow$ partially present in literature
- for small area poverty indexes $\rightarrow$ partially present in recent literature
- for small area inequality indicators $\rightarrow$ research in progress

Part II

M-quantile Estimator for Small Area Poverty and Inequality Indexes

Notation

- D: is the number of small areas of interest
- d: is the subscript for the small areas, $d = 1, \ldots, D$
- i: is the subscript for the units in a small area, $i = 1, \ldots, n_d$
- n_d: is the sample size in area d
- n: is the total sample size, $\sum_{d=1}^{d} n_d = n$
- N is the population size, while N_d is the population size in area d
- s: is the set of the sampled units and s_d is the set of sampled units in area d
- r: is the set of the non sampled units and r_d is the set of non sampled units in area d
- y_{id}: is the study variable for the unit i in the area d
- $\mathbf{x}_{id}$: is the vector of the p auxiliary variables for the unit i in area d (this vector is known for all the units in the population)

M-quantile models

- With regression models we model the mean of the variable of interest (y) given the covariates ($\mathbf{x}$)
- A more complete picture is offered, however, by modeling not only the mean of (y) given ($\mathbf{x}$) but also other quantiles. Examples include the median, the 25th, 75th percentiles. This is known as quantile regression
- The M-quantile regression model is a general framework that includes quantile regression as particular case. For a quantile q

$$Q_q = \mathbf{x}_{id}^T \boldsymbol{\beta}_\psi(q)$$

where ψ is an influence function (i.e. Huber 2 proposal)

Using M-quantile models to measure area effects

Central Idea: Area effects can be described by estimating an area specific q value ($\hat{\theta}_d$) for each area (group) of a hierarchical dataset (Chambers & Tzavidis 2006)

- Estimate the area specific target parameter by fitting an M-quantile model for each area at $\hat{\theta}_d$

$$y_{id} = \mathbf{x}_{id}^T \hat{\beta}_\psi(\hat{\theta}_d) + e_{id}$$

- Main features of these models
  - No hypothesis of normal distribution of the residuals
  - Robust methods (influence function of the M-quantile regression)

Poverty Indexes

Foster et al. (1984) define a measure of poverty based on the poverty line t and on a welfare variable y. For N units their poverty measure is

$$Z(\alpha, t) = \sum_{i=1}^{N} \left(\frac{t - y_i}{t} \right)^{\alpha} I(y_i \leq t) \quad i = 1, \ldots, N \, .$$

- $\alpha = 0$ defines the Head Count Ratio. *Incidence* of the poverty
- $\alpha = 1$ defines the Poverty Gap. *Intensity* of the poverty
- $\alpha = 2$ defines the Poverty Severity. Identify areas with severe level of poverty
- The poverty line t is generally computed as $0.6 \times \mathit{median}(y)$ and in this presentation is treated as a known value

M-quantile Poverty Mapping

Denoting by t the poverty line and by y a measure of welfare, the Foster et al. (1984) poverty measures (FGT) for a small area d can be defined as

$$Z_d(\alpha, t) = N_d^{-1}\Big[\sum_{i \in s_d} z_{id}(\alpha, t) + \sum_{k \in r_d} z_{kd}(\alpha, t)\Big]$$

where for a generic unit i in area d

$$z_{id}(\alpha, t) = \left(\frac{t - y_{id}}{t}\right)^{\alpha} \mathrm{I}(y_{id} \leqslant t) \quad i = 1, \ldots, N_d$$

- $z_{id}(\alpha, t)$ is known for $i \in s_d$
- $z_{kd}(\alpha, t)$ is unknown for $k \in r_d$ and should be predicted

Poverty Measures Estimator

Using a *smearing-type* predictor we can predict the $z_{kd}(\alpha, t)$ values

$$\hat{z}_{kd}(\alpha, t) = n_d^{-1} \sum_{i \in s_d} \left(\frac{t - \hat{y}_{ikd}}{t} \right)^{\alpha} \mathrm{I}(\hat{y}_{ikd} \leqslant t) \quad k \in r_d, i \in s_d$$

- $\hat{y}_{kd} = \mathbf{x}_{kd}^T \hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_d) \rightarrow \hat{y}_{ikd} = \mathbf{x}_{kd}^T \hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_d) + e_{id}$
- $e_{id} = y_{id} - \mathbf{x}_{id}^T \hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_d)$

The small area estimator of FGT is as follow

$$\hat{Z}_d(\alpha, t) = N_d^{-1} \left[\sum_{i \in s_d} z_{id}(\alpha, t) + \sum_{k \in r_d} \hat{z}_{kd}(\alpha, t) \right]$$

The Gini coefficient

- Economic inequality measures the disparity between a percentage of population and the percentage of resources (such as income) received by that population
- Inequality increases as disparity increases
- The Gini coefficient measures the inequality among values of a frequency distribution
- A Gini coefficient of zero expresses perfect equality, where all values are the same (for example, where everyone has an exactly equal income). A Gini coefficient of one expresses maximal inequality among values (for example where only one person has all the income)

The Gini coefficient

The Gini coefficient can be defined as follow

$$G = \left(N \sum_{i \in \Omega} y_i \right)^{-1} \left(2 \sum_{i \in \Omega} y_{(i)} i \right) - (N+1)/N,$$

where

- y_i is the income of the unit i with $y_i \geq 0$
- N is the population size
- $\Omega = \{1, \ldots, N\}$ is the set of all the units in the population
- $y_{(i)}$ are the y_i sorted in ascending order

The Gini coefficient estimator

A smearing type estimator of the Gini coefficient in area d, $\hat{G}_d$, is as follow

$$\hat{G}_d = N_d^{-1} \sum_{k \in \Omega_d} \left\{ \frac{2 \sum_{i \in s_d} \left((\hat{y}_{(k)d} + e_{id}) i \right)}{n_d \sum_{i \in s_d} (\hat{y}_{kd} + e_{id})} - \frac{n_d + 1}{n_d} \right\},$$

- $\hat{y}_{kd} = \mathbf{x}_{kd}^T \hat{\beta}_\psi(\hat{\theta}_d)$
- $\hat{y}_{(k)d}$ are the $\hat{y}_{kd}$ sorted in ascending order

REMARK: Sampled y_{kd} values $(k \in s_d)$ can be used instead of $\hat{y}_{kd}$ values

The Theil inequality index

- The Theil index is often used because it has several properties and it can incorporate group-level data
- The Theil index allows to decompose inequality into within groups and between groups components
- The Theil index is a special case of the General Entropy index $(\alpha = 1)$
- The Theil index varies between 0 and $\ln N$

The Theil inequality index

The Theil index can be defined as

$$T = \frac{V}{\mu} - \ln \mu$$

where

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad V = \frac{1}{N}\sum_{i=1}^{N} y_i \ln y_i$$

y_i is the income of person i ($y_i > 0$) and N is the number of units in the population.

The Theil index estimator

A smearing type plug-in estimator of the Theil index in area d, $\hat{T}_d$, is

$$\hat{T}_d = \frac{\hat{V}_d}{\hat{\mu}_d} - \ln \hat{\mu}_d \, ,$$

with

$$\hat{\mu}_d = N_d^{-1} \Big\{ \sum_{j \in s_d} y_{jd} + \sum_{k \in r_d} \hat{y}_{kd} + (N_d/n_d - 1) \sum_{j \in s_d} e_{jd} \Big\}$$

$$\hat{V}_d = N_d^{-1} \Big\{ \sum_{j \in s_d} y_{jd} \ln y_{jd} + n_d^{-1} \sum_{j \in s_d} \sum_{k \in r_d} (\hat{y}_{kd} + e_{jd}) \ln(\hat{y}_{kd} + e_{jd}) \Big\} ,$$

- $\hat{y}_{kd} = \mathbf{x}_{kd}^T \hat{\beta}_\psi(\hat{\theta}_d)$

Part III

MSE Estimation of Small Area Poverty and Inequality Indexes

A mean squared error estimator of the small area target estimator

To estimate the mean squared error of the M-quantile target estimator we can use the bootstrap proposed by Marchetti et al. (2012).

- Let $b = (1, \dots, B)$, where B is the number of bootstrap populations
- Let $r = (1, \dots, R)$, where R is the number of bootstrap samples
- Let $\mathbf{U} = (y_k, \mathbf{x}_k)$, $k \in (1, \dots, N)$, be the target population
- By $\cdot^*$ we denote bootstrap quantities
- $\hat{\tau}_d$ denotes the target statistic estimator of the small area d
- 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 $s = (1, \dots, n)$ be a within area simple random sample of the finite population $\Omega = \{1, \dots, N\}$

A mean squared error estimator of the small area target estimator

- Fit the M-quantile regression model on sample s, $\hat{y}_{jd} = \mathbf{x}_{jd}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d)$
- Compute the residuals, $y_{jd} - \hat{y}_{jd} = e_{jd}$
- Generate B bootstrap populations of dimension N, $\mathbf{U}^{*b} = \{y_k^*, \mathbf{x}_k\}$
    1. $y_{kd}^* = \mathbf{x}_{kd}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) + e_{kd}^*$, $k = (1, \dots, N)$
    2. e_{kd}^* are obtained by sampling with replacement residuals e_{jd}
    3. residuals can be sampled from the empirical distribution function or from a smoothed distribution function
    4. we can consider all the residuals $(e_j, j = 1, \dots, n)$, that is the unconditional approach or only area residuals $(e_{jd}, j = 1, \dots, n_d)$, that is the conditional approach.
- From every bootstrap population draw R samples of size n without replacement

A mean squared error estimator of the small area target estimator

From the B bootstrap populations and from the R samples drawn from every bootstrap population estimate the mean squared error of the target estimator as

$$\hat{E}\left[\hat{\tau}^* - \tau^*\right] = B^{-1} \sum_{b=1}^{B} R^{-1} \sum_{r=1}^{R} \left(\hat{\tau}^{*br} - \tau^{*b}\right) \quad \text{Bias}$$

$$\widehat{Var}\left[\hat{\tau}^* - \tau^*\right] = B^{-1} \sum_{b=1}^{B} R^{-1} \sum_{r=1}^{R} \left(\hat{\tau}^{*br} - \hat{\hat{\tau}}^{*br}\right)^2 \quad \text{Variance}$$

- τ^{*b} is the target statistics of the bth bootstrap population
- $\hat{\tau}^{*br}$ is the target statistics estimate of τ^{*b} estimated using the rth sample drown from the bth bootstrap population
- $\hat{\hat{\tau}}^{*br} = R^{-1} \sum_{r=1}^{R} \hat{\tau}^{*br}$

Part IV

A unified Monte Carlo Method for Poverty and Inequality Indexes

Monte Carlo M-quantile Estimators

1. Fit the M-quantile small area model using the sample values $(y_i, i = 1, \ldots, n)$ and obtain model parameters estimates β and θ_d
2. Draw an out of sample vector using

$$y^*_{id,h} = \mathbf{x}_{id,h}\hat{\beta}(\hat{\theta}_d) + e^*_{id,h}$$

   - $e^*_{id,h}$, $i = n+1, \ldots, N_d - n_d, d = 1, \ldots, D$ is drawn from the Empirical (or Smooth) Distribution Function of the M-quantile regression residuals $e_{id}, i = 1, \ldots, n_d; d = 1 \ldots, D$
   - $\hat{\beta}$, $\hat{\theta}_d$ are parameters estimates obtained from the previous step
3. Repeat the process H times. Each time combine the sample data and out of sample data for estimating the target
4. Average the results over H simulations

Monte Carlo M-quantile Estimators

The proposed Monte Carlo Estimator has the following characteristic

- It mimics the behavior of the smearing-type estimators we showed
- It is easy to implement for every statistics
- It saves memory space and it isn't too much time consuming
- The bootstrap scheme proposed works very well also with this estimator

Part V

Concluding remarks

Ongoing research

Ongoing and future research

- Estimate BES indicators at a local level
  - Point estimates
  - Root mean squared error estimates (confidence interval)
- Develop a smearing and MC estimator for the Quintile Share Ratio indicator (BES and Laeken inequality indicator)
- Develop an analytic estimator of the variance of the Theil estimator