Estimating inequality at local level in Italy

Stefano Marchetti[1,2]
[1]University of Pisa, Pisa, ITALY
[2]Corresponding author: Stefano Marchetti, e-mail: stefano.marchetti@ec.unipi.it

Abstract

Available data to compute inequality indicators in Italy come mainly from sample surveys, such as the Survey on Income and Living Conditions (EU-SILC). However, these data can be used to produce accurate estimates only at national or regional level (NUTS 2 level). To obtain estimates referring to smaller unplanned domains small area methodologies can be used. In this work I propose a smearing type estimator for the Gini's coefficient and the Theil's index to obtain estimates in the Provinces of the Tuscany Region (LAU 1 level). The proposed estimators are based on the M-quantile models, which do not impose strong distributional assumptions and are outlier robust. The use of these models for poverty and inequality estimation may protect against departures from assumptions of the traditional unit-level nested error regression model for small area estimation. In this work I also propose a model-based simulation to show the performance of the proposed estimators. Moreover, some advices on bootstrap estimation of mean squared error are given.

Keywords : Small Area, M-quantile, Robust Estimator

1 Introduction

Sample surveys provide an effective way of obtaining estimates for population characteristics. Often, in some unplanned areas or domains of interest, direct estimators, i.e. estimation based only on the sample data from the domain (Rao, 2003), are not sufficiently precise. These areas are identified with the term "small areas" and there is need of alternative methods to obtain reliable estimates, such as model-based methods.

In this paper I focus on the small area estimation of Theil and Gini inequality indexes. This work is motivated by the fact that small area estimates of inequality is yet unexplored.

In section 2 I briefly review the M-quantile small area model and present point estimation of Theil's and Gini's inequality indexes. In this section I also give some advice on mean squared error estimation using a non-parametric bootstrap approach. In section 3 the performance of the proposed estimators is empirically evaluated with model based Monte Carlo simulations. Finally, in section 4 I present estimates of Theil and Gini indexes at provincial level (LAU 1) in Tuscany.

2 Small area estimation of inequality using the M-quantile approach

In what follows I assume that a vector of p auxiliary variable $\mathbf{x}_{ij}$ is known for each population unit i in small area $j = 1, \ldots, m$ and that values of the variable of

interest y are available from a random sample, s, that includes units from all the small areas of interest. I denote the population size, sample size, sampled part of the population and non sampled part of the population in area j respectively by N_j, n_j, s_j and r_j. I assume that the sum over the areas of N_j and n_j is equal to N and n respectively. We further assume that conditional on covariate information for example, design variables, the sampling design is ignorable.

In this work I focus on the M-quantile approach to small area estimation (Chambers and Tzavidis, 2006), letting investigation on inequality estimation under the mixed models approach to future works.

Let us for the moment and for notational simplicity drop subscript j. Let $(\mathbf{x}_i^T, y_i),\ i = 1, \cdots, n$ be the observed values for a random sample of n units, where $\mathbf{x}_i^T$ are row p-vectors of known auxiliary variables and y_is are realization of a continuous random variable with unknown continuous cumulative distribution function F. The M-quantile of order q for the conditional density of y given the set of covariates $\mathbf{x}$, $f(y|\mathbf{x})$, is defined as the solution $MQ_y(q|\mathbf{x};\psi)$ of the estimating equation $\int \psi_q\{y - MQ_y(q|\mathbf{x};\psi)\}f(y|\mathbf{x})\,\mathrm{d}y = 0$. Here, ψ_q denotes an asymmetric influence function, which is the derivative of an asymmetric loss function ρ_q. A linear M-quantile regression model y_i given $\mathbf{x}_i$ is one where we assume that

$$MQ_y(q|\mathbf{x}_i;\psi) = {\mathbf{x}_i}^T\boldsymbol{\beta}_\psi(q). \tag{1}$$

Estimates of $\boldsymbol{\beta}_\psi(q)$ are obtained by minimizing

$$\sum_{i=1}^{n} \rho_q\big(y_i - {\mathbf{x}_i}^T\boldsymbol{\beta}_\psi(q)\big). \tag{2}$$

Throughout this paper I will take the linear M-quantile regression model to be defined by when ρ_q is the Huber loss function (Breckling and Chambers, 1988). Setting the first derivative of (2) equal to zero leads to the following estimating equation

$$\sum_{i=1}^{n} \psi_q(e_{iq})\mathbf{x}_i = \mathbf{0}, \tag{3}$$

where $e_{iq} = y_i - \mathbf{x}_i^T\boldsymbol{\beta}_\psi(q)$, $\psi_q(e_{iq}) = 2\psi(s^{-1}e_{iq})\{q\,I(e_{iq} > 0) + (1-q)\,I(e_{iq} \le 0)\}$ and $s > 0$ is a suitable estimate of scale, $s = (\text{median}\,|e_{iq}|)/0.6745$. I use the Huber Proposal 2 influence function, $\psi(u) = uI(-c \le u \le c) + c \cdot sgn(u)$. Provided that the tuning constant c is strictly greater than zero, estimates of $\boldsymbol{\beta}_\psi(q)$ are obtained using iterative weighted least squares (IWLS).

Using M-quantile it is possible to characterize the conditional variability across the population of interest by the M-quantile coefficients of the population units. For unit i with values y_i and $\mathbf{x}_i$, this coefficient is the value q_i such that $MQ_y(q_i|\mathbf{x}_i;\psi) = y_i$. The M-quantile coefficients are determined at the population level. Consequently, if a hierarchical structure does explain part of the variability in the population data, then we expect units within areas (or domains) defined by this hierarchy to have similar M-quantile coefficients. When the conditional M-quantiles are assumed to follow the linear model (1), with $\boldsymbol{\beta}_\psi(q)$ a sufficiently smooth function of q, the M-quantile small area model is

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}_\psi(\theta_j) + \varepsilon_{ij}\ ,$$

where θ_j is the average value of the M-quantile coefficients (q_{ij}) of the units in area j and ε_{ij} is an unit level error with distribution function G. Giving that only sampled observation are known the parameters $\boldsymbol{\beta}_\psi$ and θ_j have to be estimated.

$\boldsymbol{\beta}_\psi$ is estimated following (3) while θ_j is estimated by the sample mean of the M-quantile coefficients of sampled units in area j.

In this paper the parameters of interest are Gini and Theil inequality indexes. Given the income variable $y \geq 0$, the Gini index in area j is defined as

$$G_j = \Big(N_j \sum_{i\in\Omega_j} y_{ij}\Big)^{-1}\Big(2\sum_{i\in\Omega_j} y_{(i)j} i\Big) - (N_j+1)/N_j, \tag{4}$$

where $y_{(i)j}$ are the y_{ij} sorted in ascending order and Ω_j is the set of all the unit in area j.

A first attempt to estimate small area Gini inequality index, equation (4), is based on a smearing estimator

$$\hat{G}_j^{sm} = N_j^{-1} \sum_{k\in\Omega_j} \left\{ \frac{2\sum_{i\in s_j}\big((\hat{y}_{(k)j} + e_{ij})i\big)}{n_j \sum_{i\in s_j}(\hat{y}_{kj} + e_{ij})} - \frac{n_j+1}{n_j} \right\}, \tag{5}$$

where $\hat{y}_{kj} = \mathbf{x}_{kj}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$ with $k \in r_j$, $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$ with $i \in s_j$ and $\hat{y}_{(k)j}, k \in r_j$ are the $\hat{y}_{kj}$ sorted in ascending order. Here we suppose that link between sampled values and population values is not possible so that sampled values are used only to estimate model parameters and to compute model residuals e_{ij}.

The Theil index (that is a special case of the general entropy index) in area j can be defined as

$$T_j = \frac{\int_{-\infty}^{+\infty} y \log y \,\mathrm{d}F_j(y)}{\int_{-\infty}^{+\infty} y \,\mathrm{d}F_j(y)} - \log\Big(\int_{-\infty}^{+\infty} y \,\mathrm{d}F_j(y)\Big). \tag{6}$$

Using the smearing estimator of the cumulative distribution function proposed by Chambers and Dunstan (1986) we obtain a small area estimator for the Theil index (6) based on the M-quantile model:

$$\hat{T}_j^{sm} = \frac{\int_{-\infty}^{+\infty} t \log t \,\mathrm{d}\hat{F}_j(t)}{\int_{-\infty}^{+\infty} t \,\mathrm{d}\hat{F}_j(t)} - \log\Big(\int_{-\infty}^{+\infty} t \,\mathrm{d}\hat{F}_j(t)\Big), \tag{7}$$

where

$$\hat{F}_j(t) = N_j^{-1}\Big\{\sum_{i\in s_j} I(y_{ij} \leq t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I(\hat{y}_{kj} + e_{ij} \leq t)\Big\}.$$

Noting that

$$\int_{-\infty}^{+\infty} g(t)\,\mathrm{d}\hat{F}_j(t) = N_j^{-1}\Big\{\sum_{i\in s_j} g(y_{ij}) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} g(\hat{y}_{kj} + e_{ij})\Big\}$$

the proposed estimators (7) of the small area Theil index reduces to

$$\hat{T}_j^{sm} = \frac{N_j^{-1}\Big\{\sum_{i\in s_j} y_{ij}\log y_{ij} + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j}(\hat{y}_{kj}+e_{ij})\log(\hat{y}_{kj}+e_{ij})\Big\}}{N_j^{-1}\Big\{\sum_{i\in s_j} y_{ij} + \sum_{k\in r_j}\hat{y}_{kj} + (N_j/n_j - 1)\sum_{i\in s_j} e_{ij}\Big\}} - \log\Big(N_j^{-1}\Big\{\sum_{i\in s_j} y_{ij} + \sum_{k\in r_j}\hat{y}_{kj} + (N_j/n_j-1)\sum_{i\in s_j} e_{ij}\Big\}\Big). \tag{8}$$

When the linkage between sampled values and population values is not possible the terms $\sum_{i\in s_j} y_{ij}$ and $\sum_{i\in s_j} y_{ij}\log y_{ij}$ should be dropped from (8). This is

motivated from the fact that income y is predicted for all the population units in the area, also for the sampled ones.

Theil index is scale invariant and it can be shown that $T_j/\log N_j \in [0,1]$.

The non-parametric bootstrap scheme proposed by Marchetti et al. (2012) can be used to estimate the mean squared error of the estimators (5) and (8), but further investigation is needed. As an alternative an analytic mean squared error estimator for the Theil index is under study.

3 Empirical evaluation of performance of small area inequality estimators

In this section I use model-based Monte-Carlo simulations to empirically evaluate the performance of the small area estimator (5) and (8). The behavior of this two estimators is assessed under two scenarios for the area-specific sample and population sizes.

Population data $\Omega = (x, y)$ in $m = 30$ small areas are generated by using a unit level area random effects model with normally distributed random area effects and unit level errors as follows

$$y_{ij} = 3000 - 150 * x_{ij} + \gamma_j + \varepsilon_{ij},$$

where the area random effects $\gamma_j \sim N(0, 200^2)$, the unit level errors $\varepsilon_{ij} \sim N(0, 800^2)$, the auxiliary variable $x_{ij} \sim N(\mu_j, 1)$ where $\mu_j \sim U[4, 10]$ and μ_j was held fixed over simulations.

For each Monte Carlo simulation a within small areas random sample is selected. Two scenarios for the population and sample sizes are investigated. Under the first scenario (denoted in the tables of results by $\lambda = 1$) the total population size is $N = 8400$ with small area-specific population sizes ranging between $150 \leq N_j \leq 440$. The total sample size is $n = 840$ and the area-specific sample sizes are ranging between $15 \leq n_j \leq 44$. Under the second scenario (denoted in the tables of results by $\lambda = 2$) the total population size is $N = 2820$ with area-specific population sizes ranging between $50 \leq N_j \leq 150$ and the total sample size is $n = 282$ with area-specific sample sizes ranging between $5 \leq n_j \leq 15$. Using these two scenarios enables to assess the effect of the domain sample sizes both on the bias and the stability of the estimators.

I evaluated the performance of the proposed estimators in terms of bias, absolute bais and root mean squared error:

$$B(\hat{Z}_j) = H^{-1} \sum_{h=1}^{H} (\hat{Z}_j^h - Z_j^h) \qquad A(\hat{Z}_j) = H^{-1} \sum_{h=1}^{H} |\hat{Z}_j^h - Z_j^h|$$

$$R(\hat{Z}_j) = \left(H^{-1} \sum_{h=1}^{H} (\hat{Z}_j^h - Z_j^h)^2 \right)^{1/2},$$

where Z_j^h is the true, i.e. computed on all the population units, value of a given statistics in area j in the Monte Carlo iteration h, while $\hat{Z}_j^h$ is its estimate. So Z_j can be one of G_j or T_j. H is equal 1000 and it is the number of Monte Carlo simulations. Results are summarized over simulations and averaged over areas, they are shown in table 1.

The two estimators show a similar behavior in terms of bias, absolute bias and root mean squared error. However the Theil estimator (8) is a little bit more

	Gini		Theil	
	$\lambda = 1$	$\lambda = 2$	$\lambda = 1$	$\lambda = 2$
Bias	−0.007	−0.017	−0.001	−0.003
Abs Bias	0.016	0.033	0.021	0.029
RMSE	0.020	0.040	0.026	0.035

Table 1: Averages over areas and simulations of the bias, absolute bias and empirical (Monte Carlo) root mean squared error (RMSE) for M-quantile estimators of small area Gini and Theil inequality indexes

accurate and precise than the Gini estimator (5). This is true in the two sample size scenario. We can also see that the root mean squared error, the bias and absolute bias are bigger for the scenario with smaller sample size for both the indexes as expected.

4 Estimating inequality for provinces in Tuscany

Inequality estimation based on Gini and Theil indexes is performed by using data from the 2008 European Survey on Income and Living Conditions (EU-SILC) in Italy and the 2001 Census microdata for the region of Tuscany. In Italy provinces (LAU 1) are partitions of a region (NUTS 2) and are, with respect to EU-SILC, unplanned domains. The sample sizes for provinces in Tuscany range from 65 households in the Grosseto province to 415 households in the Florence Province with an average sample size of 149 households (median 129 households). The population of households in the different provinces based on 2001 Census data ranges from 80,810 households in the province of Massa-Carrara to 376,300 in the province of Florence with the total number of households in Tuscany being 1,388,252. Although the 2008 EU-SILC data were collected six years after the census (2008 EU-SILC data refers to 2007), the 2001–2007 period was one of relatively slow growth and low inflation in Italy, so it is reasonable to assume that there was relatively little change. These data were provided by the Italian National Institute of Statistics to the researchers of the SAMPLE project (Small Area Methods for Poverty and Living Condition Estimates, is a research project funded by the European Commission under its Seventh Framework Programme) and were analyzed by respecting all confidentiality restrictions.

Using the Tuscany EU-SILC survey data I estimated a small area working model. The response variable is the household equivalised income. The explanatory variables I considered are those that are common to the survey and Census datasets: gender, age, occupational status and years in education (variables referred to the head of the household), ownership status of the house and the number of household members. The fit of a two-level (households within provinces) random effects model using the above explanatory variables reveals departures from the assumed normality of the level 1 and level 2 error terms. For this reason, I decided to use an outlier robust model, the M-quantile small area model. The results are summarized in tables 2

According to both the indexes, inequality is higher in Grosseto, Pistoia and Arezzo. However, the level of inequality in the Tuscany provinces is lower than the inequality in Italy that has an estimated Gini coefficent in 2008 equal to 0.31 (Theil index is not available for Italy).

Further research includes alternative estimators for the Gini coefficient and

Province	Gini	Theil
Massa–Carrara	0.283	0.104
Lucca	0.284	0.086
Pistoia	0.302	0.142
Firenze	0.288	0.097
Livorno	0.247	0.082
Pisa	0.265	0.103
Arezzo	0.303	0.141
Siena	0.272	0.116
Grosseto	0.306	0.138
Prato	0.253	0.074

Table 2: Estimates of the Gini and Theil inequality indexes for provinces in Tuscany

bootstrap estimators for the mean squared error of Gini and Theil estimators, which enables for accuracy of the estimates and allows cross-sectional comparisons. Also an analytic mean squared error estimator of the Thail index is under study. Estimates has been computed by the R language (R Development Core Team, 2010).

References

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

Chambers, R. and Dunstan (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(2):255–68.

Marchetti, S., Tzavidis, N., and Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for m-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics and Data Analysis*, 56(10):2889–2902.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. (2003). *Small Area Estimation*. New York:Wiley.