Estimating the Gini concentration coefficient for the income distribution in small areas

Enrico Fabrizi [1]
Carlo Trivisano [2]

RSA European Conference 2015

Piacenza, 05/25/2015

[1]Università Cattolica del S. Cuore, Piacenza, Italy
[2]Università di Bologna, Italy

Outline

- the Gini concentration coefficient;
- small area estimation;
- estimating the Gini coefficient in small areas: current methodology;
- our proposal:
  - *area level* modelling;
  - Bayesian Beta regression;
- empirical application using EU-SILC data.

Gini concentration coefficient

- The Gini coefficient is a very popular measure for the analysis of economic inequality within a population;
- It can be defined as

$$\gamma = 2\int_0^1 (y - L(y))dy = 1 - 2\int_0^1 L(y)dy = \frac{1}{2\mu}\Delta$$

where Y is a (positive) size variable, $F(y)$ the CDF and $f(y)$ the density,
$L(y) = \mu^{-1}\int_0^{+\infty} F^{-1}(t)dt$ with $\mu = \int_0^{+\infty} yf(y)dy$ the Lorenz curve,
Δ the absolute mean difference.

Estimation of the Gini coefficient from survey data

- We are interested in estimating the Gini coefficient γ_d for subsets of the population U that we denote as U_d, $d = 1, \dots, D$;
- Available sample data: y_{dj} $(d = 1, \dots, D;\ j = 1, \dots, n_d)$;
- We assume that the sampling design is complex so that a weight w_{dj} is associated to each individual in the sample, accounting for both inequal selection probability and re-weighting adjustments for non-response;
- A survey-weighted asymptotically unbiased estimator of γ_d is given by:

$$g_d = \frac{2 \sum_{j=1}^{n_d} \left(w_{dj} y_{dj} \sum_{h=1}^{j} w_{dh} \right) - \sum_{j=1}^{n_d} y_{dj} w_{dj}^2}{\left(\sum_{j=1}^{n_d} w_{dj} \right) \left(\sum_{j=1}^{n_d} y_{dj} w_{dj} \right)} - 1$$

Small area estimation: the problem

- Large social sample surveys, such as the EU-SILC are designed to provide estimates of economic, well-being and social exclusion indicators for whole countries or large regions, social groups within countries;
- Most of these measures are often needed for a collection of geographically small areas, as indicators may be distributed unevenly among the subsets of relatively small regions;
- Often, for these small areas the available samples are not large enough to allow the ordinary survey sampling estimators to reliable;

Survey-weighted estimator g_d based on small samples

- the variance $V(g_d)$ becomes unacceptably large;
- g_d can be severely biased in small samples;

- We can see this using a simulation exercise based on synthetic the data set `eusilcP` from the R package `simFrame` (Alfons et al., 2010) generated from the real Austrian sample of the EU-SILC survey.
- In the MC experiment, we draw stratified cluster random sampling from the 9 federal states sub-samples, using households as clusters. The overall sample size in terms of households $m = 130$ allocated to strata almost proportionally.

Simulation results

Area	m_d	γ_d	$rbias(g_d)$	$rrmse(g_d)$
Burgenland	4	25.09	-0.33	0.49
Vorarlberg	6	27.85	-0.24	0.37
Salzburg	8	31.71	-0.19	0.35
Carinthia	10	26.44	-0.16	0.32
Tyrol	12	25.18	-0.12	0.31
Styria	15	25.82	-0.12	0.26
Upper Austria	20	25.55	-0.08	0.24
Lower Austria	25	25.05	-0.06	0.23
Vienna	30	29.68	-0.06	0.18

- $rbias(g_d) = B(g_d)/\gamma_d$, $rrmse(g_d) = \sqrt{MSE(g_d)}/\gamma_d$;
- We also studied the distribution of the squared income (higher concentration). The relative bias of g_d gets higher.

Small area estimation of γ_d: current methodology

In small area estimation we study how to obtain reliable estimates when domain-specific samples sizes are too small. The idea is that of complementing survey data and auxiliary information.

The 'World Bank' methodology

- estimating an econometric model for income at the household level using data from an household survey sample;
- use the estimated parameters to simulate the whole distribution from a larger data set, typically a population Census.
- calculate the Gini coefficient from these simulated data.

This methodology is due to Elbers et al. (2003) and applied in several papers and reports from the World Bank.

Possible limitations of the WB methodology

A detailed discussion of the assumptions underlying the WB methodology can be found in Tarozzi and Deaton (2009). With reference with statistical estimation we note that:

- the implementation of the method requires that information from the Census is available at the household level;
- the same vector of covariates should be available from both the survey and the Census and their measurement in the two occasions must be consistent;
- for the analysis to be meaningful, the Census and the survey year should be the same or close.

Small area estimation: area level approach

- The area-level approach is based on the idea of complemeting survey-weighted estimators with auxiliary information available for the target areas through the use of models;
- Fay-Herriot type of models are popular:

$$\begin{aligned} \hat{\theta}_d &\sim D_1\Big([\theta_d], [V_d]\Big) \\ f(\theta_d) &\sim D_2\Big([\mathbf{x}_d^t\beta], [A]\Big) \end{aligned}$$

  where $i = 1, \ldots, m$ ranges over the set of the target areas.
- In the original formulation $D_1 \equiv D_2 \equiv N(.,.)$, $f \equiv I(.)$ but alternative assumptions are also widely used, especially in the Bayesian literature.

Reducing the bias of the direct estimator

The functioning of the model we introduced hinges on the assumption

$$E(\hat{\theta}_d|\theta_d) \cong \theta_d$$

that is, the estimator is design-unbiased or nearly unbiased. This is not the case of g_d in small samples. We introduced the modified estimator

$$\tilde{g}_d = \frac{1}{2\hat{\bar{Y}}_d} \frac{\sum_{j=1}^{n_d} \sum_{k=1}^{n_d} w_{dj} w_{dk} |y_{dj} - y_{dk}|}{\hat{N}_d^2 - \sum_{h=1}^{m_d} w_{dj}^2}$$

The denominator in the Gini formula reduces the negative bias in small samples. The correction reduces to replacing n^2 with $n(n-1)$ under SRS (see Jasso, 1978; Deltas, 2003).

Back to simulation results

Area	γ_d	m_d	$rbias(g_d)$	$rrmse(g_d)$	$rbias(\tilde{g}_d)$	$rrmse(\tilde{g}_d)$
Burgenland	4	25.09	-0.33	0.49	0.01	0.53
Vorarlberg	6	27.85	-0.24	0.37	0.00	0.37
Salzburg	8	31.71	-0.19	0.35	-0.01	0.35
Carinthia	10	26.44	-0.16	0.32	-0.01	0.32
Tyrol	12	25.18	-0.12	0.31	0.01	0.32
Styria	15	25.82	-0.12	0.26	-0.02	0.25
Upper Austria	20	25.55	-0.08	0.24	0.01	0.25
Lower Austria	25	25.05	-0.06	0.23	0.01	0.24
Vienna	30	29.68	-0.06	0.18	-0.01	0.18

A Beta regression model for the Gini coefficient

A model for the Gini index can be specified as:

$$\tilde{g}_d \sim Beta\left(\frac{2\hat{\phi}_{gd}}{1+\gamma_d} - \gamma_d, \frac{2\hat{\phi}_{gd} - \gamma_d(1+\gamma_d)}{1+\gamma_d}\frac{1-\gamma_d}{\gamma_d}\right),$$

that implies

$$\begin{aligned} E(\tilde{g}_d|\gamma_d) &= \gamma_d \\ V(\tilde{g}_d|\gamma_d) &= (2\hat{\phi}_{gd})^{-1}\{\gamma_d^2(1-\gamma_d^2)\} \end{aligned}$$

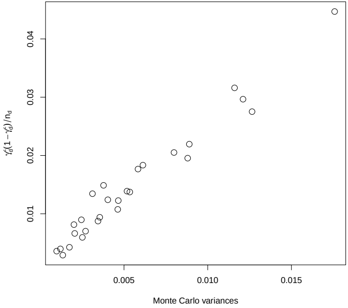
The assumption on the variance

The expression for $V(\tilde{g}_d|\gamma_d)$ can be justified in several ways:

- by assuming log-normality of y and SRS; these assumptions can be proven to lead to:

$$V_{srs}(\tilde{g}_d) \cong \frac{\gamma_d^2(1-\gamma_d^2)}{2n_d}.$$

- by simulation (the same we introduced before)



Modelling Gini coefficient: structural part

$$logit(\gamma_d) = \mathbf{x}_d^T \boldsymbol{\beta}_\gamma + v_d$$

where $\mathbf{x}_d$ contains auxiliary information for area d.

$$v_d \stackrel{ind}{\sim} N(0, \sigma_v^2)$$

For the prior of the variance component we assume:

$$\sigma_v \sim \text{half-t}(\nu = 2, A = 1)$$

(in line with Gelman, 2006).
Hyperparameters ν, A are chosen after careful consideration of the scale of the random effects and sensitivity analysis.

Estimating equivalent concentration parameters in health districts

- We have been asked to estimate several poverty related parameters
    - the at-risk-of-poverty rate,
    - the Gini coefficient,
    - the relative median at-risk-of-poverty gap,
    - material deprivation rates,

  for the health districts of the administrative region Emilia-Romagna and Tuscany.
- Health districts play a key role in the implementation of social and health expenditure programmes aimed at the contrast of social exclusion in Italy.
- Auxiliary information available for each area include average taxable income claimed by private residents, perc. of residents filling tax forms, dependency ratio, percentage of resident immigrants.

Definitions

- We use data from the EU-SILC sample survey (2010 wave);
- the Gini coefficient is based on the distribution of equivalized disposable income:

$$\text{eq.income} = \frac{\text{total disposable household income}}{\text{equivalized household size}}$$

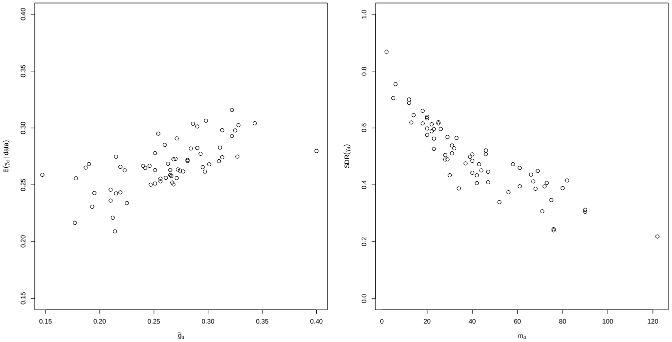
  Note that the equivalized disposable income is the same for all members of an household (i.e. we do assume 0 inequality within households);

Motivating small area methods

- Target areas: 72 health Districts;
- Population from 35.4k to 377k (115k on average);
- Overall sample size: 2692 households, 6316 individuals;
- Average sample size: 38 households, ranging from 0 (8 cases) to 253;

Survey-weighted estimators can be adequate in some cases, but they are not in most of them.

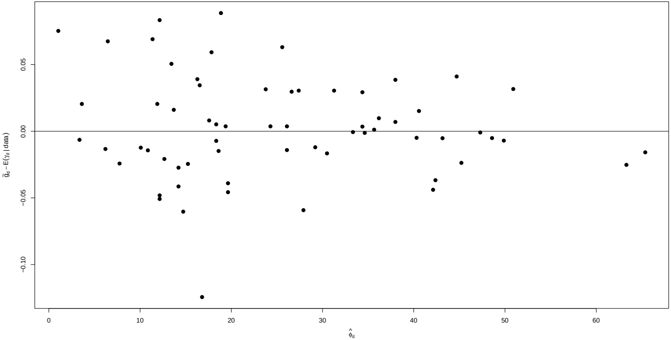
Gini coefficient: empirical results



Efficiency improvements are measured by:

$$SDR(\gamma_d) = 1 - \sqrt{\frac{V(\gamma_d|data)}{E[V(\tilde{g}_d|\gamma_d)|data]}}$$

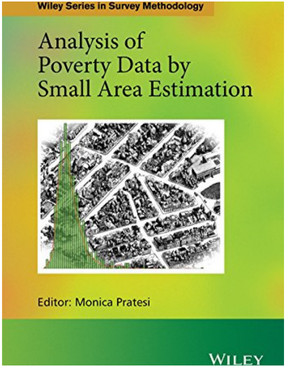
Empirical results: design consistency



Area level modelling guarantees design consistency: as the sample size gets large the small area estimator converges to the survey-weighted estimator.

References

- Fabrizi E., Trivisano C. (201?) Small area estimation of the Gini concentration coefficient, *submitted*
- E. Fabrizi, M.R. Ferrante, C. Trivisano (2015) *Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas*, in:



References

- Elbers C., Lanjouw J., Lanjouw P. (2003) Micro-level estimation of poverty and inequality, *Econometrica*, 71, 355–364.
- Griffin J.E., BrownP.J. (2010) Inference with normal-gamma prior distributions in regerssion problems, *Bayesian Analysis*, 5, 171–188;
- Jasso G. (1979), On Gini's Mean Difference and Gini's Index of Concentration, *American Sociological Review*, 44, 867-870.
- Tarozzi A., Deaton A. (2009), Using census and survey data to estimate poverty and inequality for small areas, *Review of Economics and Statistics*, 91, 773-792.