Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Small area estimation of the Gini concentration coefficient

Enrico Fabrizi [a,*], Carlo Trivisano [b]

[a] DISES, Università Cattolica del S. Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy
[b] DISES, Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna, Via Belle Arti 41, 40126 Bologna, Italy

ARTICLE INFO

Article history:
Received 29 April 2015
Received in revised form 19 January 2016
Accepted 19 January 2016
Available online xxxx

Keywords:
Hierarchical Bayes
Beta regression
Income inequality
Random effects
Markov Chain Monte Carlo
Variance components

ABSTRACT

The Gini coefficient is a popular concentration measure often used in the analysis of economic inequality. Estimates of this index for small regions may be useful to properly represent inequalities within local communities. However, the small area estimation for the Gini coefficient has not been thoroughly investigated. A method based on area level models, thereby avoiding the assumption of the availability of Census data at the micro level, is proposed. A modified design based estimator for the coefficient with reduced small sample bias is suggested as input for the small area model, while a hierarchical Beta mixed regression model is introduced to combine survey data and auxiliary information. The methodology is illustrated by means of an example based on Italian data from the European Union Survey on Income and Living Conditions.

1. Introduction

The Gini coefficient is a popular measure for the analysis of concentration. Its use for the assessment of economic inequality was popularized by Atkinson (1970), and Sen (1973) and received constant attention since then. It is routinely employed by the UN Development program and the World Bank for monitoring income inequality and, with reference to the distribution of equivalised disposable household income, is one of the set of indicators for monitoring social cohesion within the European Union by the Laeken Council (Atkinson et al., 2010).

The Gini coefficient measures concentration on a $(0, 1)$ scale, satisfies the population and Pigou–Dalton transfer principles, is scale but not translation invariant. For details, please see, for example, Cowell (2000). Many alternative concentration indexes exist (Jenkins and Van Kerm, 2009), but here we focus on the Gini coefficient for its policy use relevance in Europe.

It is important to produce estimates of income inequality for small regions to obtain a clearer picture of inequalities within these communities. In this article, we focus on the estimation of the Gini concentration coefficient within small areas, that is subpopulations for which usually the available survey samples are not large enough to provide reliable estimates. As we focus on within community concentration the non decomposability of the Gini coefficient in between and within components does not represent a relevant limitation.

The methodology for estimating the Gini coefficient from large survey samples is well established (see for instance Langel and Tillé, 2013). However, when estimating the coefficient for small geographic areas or other social groups where only a small number of samples are available, the bias of the ordinary design based estimator is non-negligible and the associated variance too large to allow for reliable inferences.

* Corresponding author.
E-mail address: enrico.fabrizi@gmail.com (E. Fabrizi).

http://dx.doi.org/10.1016/j.csda.2016.01.010
0167-9473/

The literature on estimating the Gini index for small areas is not rich, and the most notable applications are based on the methodology of Elbers et al. (2003). Example applications include Elbers et al. (2005), Demombynes and Özler (2005), and Nguyen et al. (2010). Elbers et al. (2003) method estimates an econometric model for income (or expenditure) at the household level using data from household surveys, and then uses the estimated parameters to simulate the whole distribution from a larger dataset, typically a population census. A discussion of the assumptions underlying this method and its limitations can be found in Tarozzi and Deaton (2009).

The method requires that information from the census is available at household level; the same vector of covariates should be available from the survey and census, and their measurement in two occasions must be consistent; and that for the analysis to be meaningful, the census and the survey year should be the same (or close). Tarozzi and Deaton (2009) noted that the assumed relationship between the target and the auxiliary variables should hold for the whole population apart from area or community specific random effects, which may not always be the case.

Microsimulation based small area estimation methods (see for instance Anderson, 2013; Miranti et al., 2009) similarly require a set of calibration variables whose total at the area level is known from the Census available at the micro-level, with consistent measurement, for the sample being analyzed.

We propose a different strategy that does not assume the availability of census data at the micro-level and does not require the specification of a model linking target and auxiliary variables at the household level. We combine standard survey weighted direct estimators of the Gini coefficient with auxiliary information accurately known at the area level from administrative sources, adopting an approach known as area level modeling (see Rao, 2003, ch. 5). These area level or Fay–Herriot type models rely on the assumption of approximated unbiasedness of the direct estimators, which is not even approximately satisfied by the standard survey estimators of the Gini index when based on small samples. Moreover, since the index can take values in the (0, 1) interval and is likely to have a positively skewed sampling distribution, normality based linear mixed models are not suitable.

This paper proposes a modified design based estimator of the Gini coefficient characterized by a smaller bias in small samples and equivalent to the usual survey weighted estimator for larger sample sizes. To link the Gini index estimators to area level information, we consider Beta regression models, as the Beta distribution is very flexible over the (0, 1) range and allows for asymmetric sampling distributions.

Beta regression models have received considerable attention in recent years in both frequentist and Bayesian literature. We adopt the Bayesian approach to estimation, and some basic Refs. (Branscum et al., 2007; Da Silva et al., 2011; Bayes et al., 2012). Figueroa-Zuñiga et al. (2013) discuss Bayesian mixed Beta regression; whereas Liu et al. (2014); Fabrizi et al. (2011) apply Beta regression to small area estimation of proportions.

The methodology we propose is illustrated by an empirical exercise where we estimate the Gini coefficient for the distribution of the equivalised disposable household income in small geographic regions (health districts) within two administrative regions (Emilia-Romagna and Tuscany) located in Central Italy.

Health districts are relevant territorial units for the implementation of social cohesion policies and apportioning of related funds. Our survey data source is the Italian 2010 sample of the European Union Survey on Income and Living Conditions (EU-SILC). See Fusco et al. (2010) for detailed definitions of equivalized disposable income and related concepts in the context of the EU-SILC. The population of the health districts ranges between 35.4 and 377 thousand, and they play an important role in the implementation of many social and health expenditure programmes related to the contrast of poverty and social exclusion, for which regional governments are responsible. The auxiliary information that we use is based on publicly available archives at the municipal level, and we aggregated information to obtain district level variables.

We introduce the Gini concentration coefficient and its estimation from complex survey samples in Section 2, along with the proposal for reducing bias in small samples. We also discuss the sensitivity of the Gini index estimators to outliers and a possible remedy based on Pareto tail modeling (Alfons et al., 2013). The Beta regression model for small area estimation is described in Section 3. The application to a real dataset is introduced and results are outlined in Section 4, and Section 5 provides some concluding remarks.

2. Gini concentration coefficient for the distribution of equivalized household income

The Gini coefficient is a popular inequality measure that can be applied to the distribution of income, consumption, or any other size variable defined over a population of statistical units. Its definition is related to the Lorenz curve, a function that maps the cumulative proportion of ordered individuals onto the corresponding cumulative distribution of the size variable. If Y denotes this (positive) size variable, $F(y)$ the cumulative distribution, and $f(y)$ the associated density, the Lorenz curve is $L(u) = \mu^{-1}\int_0^u F^{-1}(t)dt$ with $\mu = \int_0^{+\infty} yf(y)dy$. The Gini coefficient can be defined as

$$\gamma = 2\int_0^1 (u - L(u))du = 1 - 2\int_0^1 L(u)du.$$

There are several equivalent definitions of γ (Yitzhaki, 2002), including

$$\gamma = \frac{1}{2\mu}\Delta, \qquad (1)$$