Estimation of Poverty Measures in Small Areas

Isabel Molina[1] and J. N. K. Rao[2]

[1] Departamento de Estadística, Universidad Carlos III de Madrid, isabel.molina@uc3m.es

[2] School of Mathematics and Statistics, Carleton University, jrao@math.carleton.ca

Abstract

The eradication of extreme poverty and hunger is the first of the Millennium Development Goals established by the United Nations. The availability of reliable statistical information on living conditions of people is a basic requirement for the achievement of this goal. However, sample sizes in many national surveys are not large enough to allow reliable estimation for small areas or domains, and therefore small area estimation techniques that borrow information across areas through linking models are required. Linking models are based on auxiliary information such as census data and current administrative records. This paper deals with estimation of poverty measures in small areas. Many poverty measures are complex nonlinear functions of the values of a welfare variable for the population units. Then, the usual small area estimation methods are not directly applicable. In this paper, we propose to use empirical best predictors of the poverty measures, which are obtained through Monte Carlo simulation. Their mean squared errors are estimated using a bootstrap method. Simulation experiments are carried out both under the model-based and the design-based approaches to study the performance of the empirical best predictors.

Keywords: Parametric bootstrap; Poverty mapping; Small area estimation.

1 FGT poverty measures for small areas

Consider a finite population of size N that is partitioned into D small areas of sizes $N_1, \dots, N_D$. Let E_{dj} be a suitable quantitative measure of welfare for individual j in small area d, such as income or expenditure, and let z be the given poverty line; that is, the threshold for E_{dj} under which a person is considered as "under poverty". The family of poverty measures introduced by Foster et al. (1984) and called throughout the paper as FGT poverty measures, for each small area d, is defined as

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D,$$

where $I(E_{dj} < z) = 1$ if $E_{dj} < z$ (person under poverty) and $I(E_{dj} < z) = 0$ if $E_{dj} \geq z$ (person not under poverty). For $\alpha = 0$ we get the proportion of individuals under poverty in small area d, also called poverty incidence or head count ratio. The FGT measure for $\alpha = 1$ is called poverty gap, and measures the area mean of the relative distance to non-poverty (the poverty gap) of each individual. For $\alpha = 2$ the measure is called poverty severity. This measure squares the poverty gaps, and thus emphasizes extreme poverty.

In the inference process, a random sample of size $n < N$ is drawn from the population according to a specified sampling design. Let s be the set of units selected in the sample and r the set of indexes of the units that are not selected (with size $N - n$). The restrictions of s, N and n to area d are denoted by s_d, N_d and n_d respectively, where $n = n_1 + \cdots + n_D$. The sample FGT poverty measures are given by

$$f_{\alpha d} = \frac{1}{n_d} \sum_{j \in s_d} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D. \qquad (1)$$

A direct estimator for a small area uses only sample data from the target area and it is usually design based. Let w_{dj} be the sampling weight (inverse of the probability of inclusion) of individual j from area d. Direct estimators of the FGT measures are

$$f_{\alpha d}^{w} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_{dj} \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D, \qquad (2)$$

where $\hat{N}_d = \sum_{j \in s_d} w_{dj}$ is the direct estimator of the population size of small area d, N_d. If the sampling weights w_{dj} do not depend on the unit j, for example $w_{dj} = n_d/N_d$ under simple random sampling within areas, then (2) reduces to the sample measure (1).

The limited sample sizes n_d within some of the areas prevent the use of estimators such as (1) or (2). Indeed, a common definition of poverty classifies a person as "under poverty" when the selected welfare variable for this person is below 60% of the median. Under this definition the outcome of being under poverty is likely to have low frequency. Then, to obtain reliable estimators for small domains or geographical areas it is necessary to appeal to small area techniques (Rao, 2003). These techniques improve the estimation procedures by using models that establish some relationships between the areas, based on auxiliary information (census and/or administrative variables) related to the welfare variables of interest. These models provide "indirect" estimators that make use of related data from other areas, and which might reduce drastically the estimation errors as long as model assumptions hold.

2 Empirical best prediction of poverty measures

Suppose that there is a one-to-one transformation $Y_{dj} = T(E_{dj})$ of the welfare variables E_{dj}, such that the vector $\mathbf{y}$ containing the values of the transformed variables Y_{dj} for all population units satisfies $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{y}_s$ be the sub-vector of $\mathbf{y}$ corresponding to sample elements and $\mathbf{y}_r$ the sub-vector of out-of-sample elements, that is, $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$. Then $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ is Normal with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_s', \boldsymbol{\mu}_r')'$ and covariance matrix $\mathbf{V}$ partitioned conformably as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix},$$

and the conditional distribution of $\mathbf{y}_r$ given $\mathbf{y}_s$ is given by

$$\mathbf{y}_r|\mathbf{y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}), \tag{3}$$

where

$$\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s), \quad \mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr}. \tag{4}$$

The Best Predictor (BP) of $F_{\alpha d}$ is the function of the sample data that minimizes the mean squared error. This predictor is the conditional expectation

$$\hat{F}_{\alpha d}^{B} = E_{\mathbf{y}_r}(F_{\alpha d}|\mathbf{y}_s). \tag{5}$$

The expectation in (5) cannot be calculated explicitly using (3) and (4) because $F_{\alpha d}$ is a complex non-linear function of $\mathbf{y}$. Therefore, we propose to use an empirical approximation by Monte Carlo simulation of a large number L of vectors $\mathbf{y}_r$ generated from (3).

When the joint distribution of $\mathbf{y}$ depends on unknown parameters, an empirical BP (EBP) can be obtained by replacing all unknown parameters by suitable estimators and then using the estimated distribution to obtain a Monte Carlo approximation of (5).

Here we assume that a one-to-one transformation $Y_{dj} = T(E_{dj})$ of the welfare variables E_{dj} follows a nested error linear regression model (Battese et al., 1988), in which the variables Y_{dj} (e.g., log-earnings) are related to a vector of p explanatory variables $\mathbf{x}_{dj}$, including random area-specific effects u_d,

$$\begin{aligned} &Y_{dj} = \mathbf{x}_{dj}'\boldsymbol{\beta} + u_d + e_{dj}, \quad u_d \sim \text{iid } N(0, \sigma_u^2), \\ &e_{dj} \sim \text{iid } N(0, \sigma_e^2), \quad j = 1, \dots, N_d, \quad d = 1, \dots, D, \end{aligned} \tag{6}$$

where the area effects u_d and the errors e_{dj} are independent.

Small sample properties of EB estimators based on model (6) are analyzed by model-based and design-based simulation studies. Results show big improvements in mean squared error over direct estimators and estimators obtained by simulated censuses (Elbers et al., 2003). An application is given to estimate poverty incidences and poverty gaps in Spanish provinces by sex with mean squared errors estimated by parametric bootstrap (González-Manteiga et al., 2008). In the Spanish data, results show a non-negligible reduction in coefficient of variation of the proposed EB estimators over direct estimators for most domains.

References

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988) An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association*, 83, 28–36.

Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2003) Micro-level estimation of poverty and inequality, *Econometrica*, 71, 355–364.

Foster, J., Greer, J. & Thorbecke, E. (1984) A class of decomposable poverty measures, *Econometrica*, 52, 761–766.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. & Santamaría, L. (2008) Bootstrap Mean Squared Error of a Small-Area EBLUP, *Journal of Statistical Computation and Simulation*, 75, 443–462.

Rao, J. N. K. (2003) *Small Area Estimation*, Wiley, London.