Recent Advances in Estimation of Poverty Indicators for Domains and Small Areas

Risto Lehtonen and Ari Veijanen

Abstract In the paper, we consider the estimation of indicators on poverty and social exclusion for population subgroups or domains and small areas. For at-risk-of poverty rate, we discuss indirect estimators including model-assisted logistic generalized regression estimators and new model calibration estimators. Logistic mixed models are used in these methods. For quintile share ratio, indirect model-based synthetic estimators and new calibration-based predictor-type methods using linear mixed models are considered. Unit-level auxiliary data are incorporated in the estimation procedures. Design-based direct estimators that do not use auxiliary data and models are used for comparison. Design bias and accuracy of estimators are examined with simulation experiments using register data maintained by Statistics Finland and semi-synthetic data generated from the EU-SILC survey.

1. Introduction

There are increasing demand in Europe and elsewhere for reliable statistics on poverty and social exclusion produced for regions and other population subgroups or domains. Small area estimation of indicators on poverty and social exclusion has been recently investigated in research projects funded by European Commission under the 5th and 7th Framework Programmes (FP5 and FP7). Model-based small area estimation methods were studied in the FP5 project EURAREA (Enhancing Small Area Estimation Techniques to meet European Needs, 2002-2004). The aim of the FP7 project SAMPLE (Small Area Methods for Poverty and Living Condition Estimates, 2008-2011) was to develop new indicators for inequality and poverty with attention to social exclusion and deprivation, as well as to develop and implement methods for small area estimation of the traditional and new indicators. The FP7 project AMELI (Advanced Methodology

[1] Risto Lehtonen, University of Helsinki, email: risto.lehtonen@helsinki.fi

[2] Ari Veijanen, Statistics Finland, email: ari.veijanen@stat.fi

for European Laeken Indicators, 2008-2011) included several specialized sub-projects (work packages) and covered a wide range of topics on poverty, social exclusion and social cohesion. Conceptual background, indicator construction and measurement and estimation of indicators on poverty and social exclusion were discussed (Münnich et al., 2011). A sub-project concentrated on small area estimation methods of selected poverty indicators (Lehtonen et al., 2011).

Indicators on poverty and social exclusion investigated in AMELI included at-risk-of poverty rate, relative median at-risk-of poverty gap, quintile share ratio and the Gini coefficient. In this paper, we discuss the methods for small area estimation of poverty rate and quintile share ratio introduced in Lehtonen et al. (2011) and developed further in Veijanen and Lehtonen (2011) and Lehtonen and Veijanen (2012). Unit-level auxiliary data are incorporated in the estimation procedures. Design-based direct estimators that do not use auxiliary data and models are used for comparison. Design bias and accuracy of estimators are examined with design-based simulation experiments using register data maintained by Statistics Finland and semi-synthetic data generated from the EU-wide SILC survey (Statistics on income and living conditions).

The paper is organized as follows. Estimation for poverty rate is examined in Section 2. Methods for quintile share ratio are discussed in Section 3.

2. Estimation of poverty rate for regions

For poverty rate, we discuss indirect estimators including model-assisted logistic generalized regression estimators (Lehtonen and Veijanen, 1998; Lehtonen, Särndal and Veijanen, 2003, 2005; Lehtonen and Veijanen 2009) and new model calibration estimators (Lehtonen et al., 2011; Lehtonen and Veijanen, 2012). Logistic mixed models are used in these methods.

In classical *model-free calibration* (Deville and Särndal, 1992; Särndal, 2007), a calibration equation is imposed: the weighted sample totals of auxiliary variables reproduce the known population totals. In *model calibration* introduced by Wu and Sitter (2001), a model is first fitted to the sample. Calibration weights are determined using the fitted values instead of the original auxiliary variables: the weighted sample total of fitted values reproduces the population total of predictions. Our calibration equations for domain estimation specify that the weighted total of fitted values over a subgroup of the sample equals the sum of predictions over the corresponding population subgroup.

A model calibration procedure for domain estimation consists of two phases, the *modelling phase* and *the calibration phase*. There is much flexibility in both phases. We have chosen a mixed model formulation involving components that account for spatial heterogeneity in the population. The predictions calculated in the modelling phase are used in the calibration phase when constructing calibration equation and a calibrated domain estimator. Calibration can be defined at the population level, at the domain level or at an intermediate level, for example at a regional level (neighbourhood) that contains the domain of interest. Further, in the construction of the calibrated domain estimator, a "semi-direct" approach involves using observations only from the domain of interest, whereas in a "semi-indirect" approach, also observations outside the domain of interest are included.

The finite population is denoted $U = \{1, 2, ..., k, ..., N\}$, where *k* refers to the label of population element. A *domain* U_d is a subset of *U* such as a regional population. The number of units in the domain is denoted by N_d. In sample *s*, the corresponding subset is defined as $s_d = U_d \cap s$; it has n_d observations. The domains are of unplanned type. Inclusion probabilities are π_k and design weights are $a_k = 1/\pi_k$.

In order to account for possible differences between regions, a mixed model incorporates domain-specific random effects $u_d \sim N(0, \sigma^2_{u_d})$ for domain U_d, or regional random effects $u_r \sim N(0, \sigma^2_{u_r})$ for region U_r, where $U_d \subset U_r$. We next consider the case of u_d. For a binary *y*-variable, a logistic mixed model is of the form

$$E_m(y_k \mid u_d) = P\{y_k = 1 \mid u_d; \boldsymbol{\beta}\} = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)},$$

where $\mathbf{x}_k$ is a known vector value for every $k \in U$ and $\boldsymbol{\beta}$ is a vector of fixed effects common for all domains. The parameters $\boldsymbol{\beta}$ and $\sigma^2_{u_d}$ are first estimated from the data, and estimates $\hat{u}_d$ of the random effects u_d are then calculated. Predictions $\hat{y}_k = P\{y_k = 1 \mid \hat{u}_d; \hat{\boldsymbol{\beta}}\}$ are calculated for every $k \in U$.

The domain total of a study variable *y* is defined by

$$t_d = \sum_{k \in U_d} y_k \;, \tag{1}$$

where y_k denotes the value of the study variable for element *k*. *Horvitz-Thompson (HT) estimator* of domain total (1) is a direct estimator as it only involves observations from the domain of interest:

$$\hat{t}_d = \sum_{k \in s_d} a_k y_k . \tag{2}$$

The estimator is design unbiased but it can have large variance, especially for small domains. HT does not incorporate any auxiliary data.

Generalized regression (GREG) estimators (Särndal et al., 1992; Lehtonen and Veijanen, 2009) are assisted by a model fitted to the sample. By choosing different models we obtain a family of GREG estimators with same form but different predicted values (Lehtonen et al., 2003, 2005). Ordinary GREG estimator

$$\hat{t}_{d;GREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \tag{3}$$

incorporating a linear fixed-effects regression model is often used to estimate domain totals (1) of a continuous study variable. For a binary or polytomous response variable, a logistic model formulation is often chosen. LGREG (logistic GREG; Lehtonen and Veijanen, 1998) estimates the frequency f_d of a class *C* in each domain. A logistic regression model is fitted to indicators $v_k = I\{y_k \in C\}$, $k \in s$, using the design weights. In the MLGREG estimator (Lehtonen et al. 2005), we use a logistic mixed model for (4) involving fitted values $\hat{p}_k = P\{v_k = 1 \mid \hat{\mathbf{u}}_d; \mathbf{x}_k, \hat{\boldsymbol{\beta}}\}$. The random effects are associated with domains U_d or with larger regions U_r. The MLGREG estimator of the class frequency in U_d is

$$\hat{f}_{d;MLGREG} = \sum_{k \in U_d} \hat{p}_k + \sum_{k \in s_d} a_k (v_k - \hat{p}_k) . \tag{4}$$

The calculation of $\hat{p}_k$ for all $k \in U_d$, $d = 1,...,D$ requires access to unit-level population data on auxiliary variables.

In population level calibration (Wu and Sitter, 2001), the weights must satisfy calibration equation

$$\sum_{i \in s} w_i z_i = \sum_{i \in U} z_i = \left(N, \sum_{i \in U} \hat{y}_i \right), \tag{5}$$

where $z_i = (1, \hat{y}_i)'$. Using the technique of Lagrange multiplier (λ), we minimize

$$\sum_{k \in s} \frac{(w_k - a_k)^2}{a_k} - \lambda' \left(\sum_{i \in s} w_i z_i - \sum_{i \in U} z_i \right)$$

subject to the conditions (5). The equation is minimized by weights

$$w_k(\lambda) = a_k \left(1 + \lambda' z_k\right), \tag{6}$$

where $\lambda' = \left(\sum_{i \in U} z_i - \sum_{i \in s} a_i z_i \right)' \left(\sum_{i \in s} a_i z_i z_i' \right)^{-1}$.

In domain estimation, these weights are applied over a domain: the estimator is

$$\hat{f}_{d;pop} = \sum_{k \in s_d} w_k y_k \ . \tag{7}$$

A straightforward generalization of the population-level calibration equation is a domain-level calibration equation

$$\sum_{i \in s_d} w_{di} z_i = \sum_{i \in U_d} z_i = \left(N_d, \sum_{i \in U_d} \hat{y}_i \right), \tag{8}$$

where the weights w_{di} are specific to the domain. From (8) we see that the domain sizes must be known. We minimize

$$\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k} - \lambda_d' \left(\sum_{i \in s_d} w_{di} z_i - \sum_{i \in U_d} z_i \right)$$

subject to (8). The solution is $w_{dk} = w_k(\lambda_d)$, defined by (6) for

$$\lambda_d' = \left(\sum_{i \in U_d} z_i - \sum_{i \in s_d} a_i z_i \right)' \left(\sum_{i \in s_d} a_i z_i z_i' \right)^{-1} .$$

The domain estimator is then a weighted domain sum

$$\hat{f}_{d;s} = \sum_{k \in s_d} w_{dk} y_k \ . \tag{9}$$

We call this estimator “semi-direct”, as the sum only contains y-observations from the domain of interest. It is not a direct estimator, however, as the weights are determined by a model that is fitted to the whole sample. Various semi-direct calibration estimators are possible; see Lehtonen and Veijanen (2012).

We introduce next various new “semi-indirect” estimators. They are weighted sums over a set that is larger than the domain of interest. Our goal is to “borrow strength” from other domains, in an attempt to reduce mean squared error. A semi-indirect domain estimator incorporates whole sample, an enclosing aggregate of regions in a hierarchy of regions or the set of neighbouring domains, including the domain itself. A neighbourhood of a region comprises regions that share a common border with the specified region or regions with centre closer than a given distance threshold. In a semi-indirect estimator, we use supersets $C_d \supset U_d$ of domains with corresponding sample subsets $r_d = C_d \cap s$. In our simulations the supersets are composed of domains. We define the domain estimator as a weighted sum of all observations in r_d:

$$\hat{f}_{d;r} = \sum_{k \in r_d} w_{dk} y_k \tag{10}$$

The calibration equation is

$$\sum_{i \in r_d} w_{di} z_i = \sum_{i \in U_d} z_i \tag{11}$$

Note that the sum on the left side of (11) extends over r_d which corresponds to population subset C_d, a larger set than U_d on the right side of the equation. We have required that the weights w_{dk} are close to weights a_k in the domain and close to zero outside the domain. The weights minimize

$$\sum_{k \in r_d} \frac{(w_{dk} - I_{dk} a_k)^2}{a_k}$$

where $I_{dk} = I\{k \in s_d\}$, subject to the calibration equations (11) when

$$w_{dk} = I_{dk} a_k + \lambda_d' a_k z_k; \quad \lambda_d' = \left(\sum_{i \in U_d} z_i - \sum_{i \in r_d} I_{di} a_i z_i \right)' \left(\sum_{i \in r_d} a_i z_i z_i' \right)^{-1}.$$

Variance estimation of GREG estimators can be handled analytically (Lehtonen and Veijanen, 2009) but there is not yet theory of variance estimation of model calibration estimators for domains, so bootstrap is recommended (Gershunskaya et al., 2009).

At-risk-of-poverty rate is the proportion of poor people in a domain with equivalized income at or below the poverty line *t*. Our goal is to estimate $R_d = (1/N_d) \sum_{k \in U_d} I\{y_k \le 0.6M\}$. An estimate $\hat{M}$ of reference median income *M* is obtained from the HT estimated distribution function $\hat{F}_U(t) = (1/\hat{N}) \sum_{k \in s} a_k I\{y_k \le t\}$. The distribution function defined in domain U_d is estimated by HT: $\hat{F}_d(t) = (1/\hat{N}_d) \sum_{k \in s_d} a_k I\{y_k \le t\}$, where $\hat{N}_d = \sum_{k \in s_d} a_k$.

Direct (default) *HT-CDF estimator* of poverty rate is

$$\hat{r}_{d;HT} = \hat{F}_d(0.6\hat{M}). \tag{12}$$

To estimate domain poverty rate by MLGREG or model calibration, we first estimate the domain total of a *poverty indicator* $v_k = I\{y_k \le 0.6\hat{M}\}$, which equals 1 for persons with income below or at the poverty line and 0 for others. The estimate of the domain total t_d is then divided by the known domain size N_d (or, its estimate $\hat{N}_d$). For example, the MLGREG estimator of the poverty rate is

$$\hat{r}_{d;MLGREG} = \hat{f}_{d;MLGREG} / N_d. \tag{13}$$

For design-based simulation experiments, an artificial population of one million persons was constructed from real income data of Statistics Finland for seven NUTS level 3 regions in Western Finland. In the simulations, *K* = 1000 samples of *n* = 5000 persons were drawn with without-replacement probability proportional to size (PPS) sampling from the unit-level population. For PPS, an artificial size variable was generated as a function of the socio-economic status of household head. People with low income appear in samples with larger probability than people with large income.

Our models incorporated the following auxiliary variables: age class (0-15, 16-24, 25-49, 50-64, 65- years), gender with interactions with age class, socio-economic status of the household head (wage and salary earners, farmers, other entrepreneurs, pensioners, and others), and labour force status (employed, unemployed, and not in workforce). We created indicators for each class of a qualitative variable. As domains we used the 36 NUTS4 regions. The NUTS classification is hierarchical: each NUTS4 region is contained within a larger NUTS3 region.

The methods were nearly design unbiased by the construction principle (bias results not shown). The accuracy was measured by relative root mean squared error:

$$RRMSE = \sqrt{(1/K)\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)^2} \;/\; \theta_d \;.$$

We present the averages of RRMSE over domain classes defined by expected domain sample size (Table 1): Minor (0-50 units), Medium-sized (50-100) and Major (100-) domains. The logistic mixed model contains regional random intercepts associated with NUTS4 regions. Design weights were incorporated into the model fitting.

All methods except calibration at population level outperformed the direct estimator. In small domains, MLGREG had slightly better accuracy than calibration methods. The choice of the model did not have much effect on most estimators.

Table 1. Mean relative root mean squared error (RRMSE) (%) of poverty rate estimators over domain size classes, under logistic mixed model formulation.

Estimator	Expected domain sample size			All
	Minor	Medium	Major	
Direct	41.1	28.9	18.0	26.7
MLGREG	39.6	28.6	17.8	26.2
Semi-indirect model calibration estimators				
SI-population	39.7	28.6	17.8	26.2
SI-regional	39.7	28.5	17.8	26.2
SI-spatial	39.7	28.6	17.8	26.2

3. Estimation of quintile share ratio for regions

In an indirect model-based predictor-type estimator for quintile share ratio (QSR) based on unit-level auxiliary data, predictions obtained from a linear mixed model are plugged into the formula of QSR defined at the population level. The estimator is expected to have small variance but as a model-based estimator, it can suffer from serious design bias. To decrease design bias, we define a transformation that brings the percentiles of transformed predictions closer to the percentiles of sample values (Veijanen and Lehtonen, 2011). To account for domain differences, a linear mixed model incorporates domain-specific random effects $u_d \sim N(0, \sigma_u^2)$. The model is

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k, k \in U_d\,,\ \varepsilon_k \sim N(0,\sigma^2)\ .$$

The random effects may also be associated with aggregates of domains. The parameters $\boldsymbol{\beta}$, σ_u^2 and σ^2 are first estimated from the data by using ML or REML methods, and the values of the random effects are then predicted. This yields predictions $\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d, k \in U_d$.

QSR compares the average equivalized incomes in the poorest and the richest quintile. Each quintile contains 20 % of people; in the design-based case accounting for 20 % of design weights. The *default* (direct) *estimators* of the first (S20) and the fifth quintile average (S80) are Hájek estimators. The direct quintile share estimate is the ratio of S20 to S80. The predictor-type estimator of quintile share in a domain is the ratio of averages of predictions in the first and fifth quintiles.

Linear mixed model is fitted to $z_k = \log(y_k + 1)$ and the fitted values $\hat{z}_k$ are back-transformed to $\hat{y}_k = \exp(\hat{z}_k) - 1$. We correct for bias and spread of predictions by a nonlinear transformation that brings the distribution of predictions closer to the distribution of observed values y_k ($k \in s_d$) in terms of percentiles, denoted by $\hat{p}_{cd}$ and p_{cd} , respectively. The percentiles p_{cd} of sample values are obtained from the estimated CDF $\hat{F}_{HT;d}(y) = (1/\hat{N}_d)\sum_{k \in s_d} a_k I\{y_k \le y\}$. Our goal is to obtain transformed predictions $\tilde{y}_k = e^{\alpha_d}\hat{y}_k^{\gamma}$ whose percentiles, denoted $\tilde{p}_{cd}$, are close to p_{cd} on logarithmic scale. To avoid unstable estimates in the smallest domains, we pooled the percentile data from all domains and minimized

$$\sum_d \sum_{c=1}^{C} \left(\log\left(\tilde{p}_{cd}\right) - \log\left(p_{cd}\right)\right)^2 = \sum_d \sum_{c=1}^{C} \left(\alpha_d + \gamma \log\left(\hat{p}_{cd}\right) - \log\left(p_{cd}\right)\right)^2 ,$$

where C = 99. The *percentile-adjusted*, or *p-adjusted*, predictions involve OLS estimates of parameters α_d and γ :

$$\log\left(\tilde{y}_k\right) = \hat{\alpha}_d + \hat{\gamma}\log\left(\hat{y}_k\right) (k \in U_d)\ . \qquad (14)$$

The transformation (14) was applied only to the positive predictions, with percentiles $\hat{p}_{cd}$ and p_{cd} calculated from positive predictions and sample values.

In simulation experiments we used a semi-synthetic data set of about ten million persons, constructed from SILC data sets (Alfons et al., 2011) to mimic the regional and demographic variation of income statistics in the EU. We applied SRSWOR (n = 2000). As domains we used 40 regions. The domains were classified to minor, medium and major domains by expected sample size with class boundaries at 45 and 55 units. Our models fitted to equivalized income variable incorporated age class and gender with interactions, attained education level (ISCED), activity (working, unemployed, retired, or otherwise inactive) and degree of urbanisation of residence (three classes). The mixed models with random intercepts associated with regions, were fitted using ML. K = 1000 samples were drawn. Design bias and accuracy were assessed by absolute relative bias $ARB = |(1/K)\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)|/\theta_d$ and RRMSE (see Section 2).

Accuracy of the new p-adjusted predictor was much better than that of the default (direct) estimator, in all domain size classes (Table 2). However, the estimator was still design biased, especially in small domains, and the bias was somewhat larger than that of the direct estimator. For MSE estimation, different variants of bootstrap can be used (e.g. Gershunskaya et al., 2009).

Table 2. Results with quintile share ratio QSR in regions.

Estimator	ARB (%)				RRMSE (%)			
	Expected domain sample size				Expected domain sample size			
	Minor	Medium	Major	All	Minor	Medium	Major	All
Direct	4.9	4.6	3.4	4.4	43.5	41.7	38.5	41.3
p-adjusted predictor	12.3	8.6	5.7	8.9	16.0	13.6	11.4	13.7

References

1. Alfons, A., Templ, M., Filzmoser, P., Kraft, S., Hulliger, B, Kolb, J.-P., Münnich, R.:. Report on outcome of the simulation study. Research Project Report WP6 (D6.2, FP7-SSH-2007-217322 AMELI). Available at: http://svn.uni-trier.de/AMELI (2011)
2. Deville, J.-C., Särndal, C.-E.. Calibration estimators in survey sampling. Journal of the American Statistical Association **87**, 376–382 (1992)
3. Gershunskaya, J., Jiang, J., Lahiri, P.: Resampling methods in surveys. Chapter 28 in Rao C.R. and Pfeffermann D. (Eds.) Handbook of Statistics Vol. 29B. Sample Surveys. Inference and Analysis. Elsevier, Amsterdam, 219–249 (2009)
4. Lehtonen, R., Särndal, C.-E., Veijanen, A.: The effect of model choice in estimation for domains, including small domains. Survey Methodology Journal **29**, 33–44 (2003)
5. Lehtonen, R., Särndal, C.-E.,Veijanen, A.: Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. Statistics in Transition **7**, 649–673 (2005)
6. Lehtonen, R., Veijanen, A., Myrskylä, M., Valaste, M.: Small Area Estimation of Indicators on Poverty and Social Exclusion. Research Project Report WP2 (D2.2, FP7-SSH-2007-217322 AMELI). Available at: http://svn.uni-trier.de/AMELI (2011)
7. Lehtonen, R., Veijanen, A.: Logistic generalized regression estimators. Survey Methodology Journal **24**, 51–55 (1998)
8. Lehtonen, R., Veijanen, A.: Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.) Handbook of Statistics Vol. 29B. Sample Surveys. Inference and Analysis. Elsevier, Amsterdam 219–249 (2009)
9. Lehtonen, R., Veijanen, A.: Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics **66**, 125-133 (2012)
10. Münnich, R., Zins, S., Alfons, A., Bruch, C., Filzmoser, P., Graf, M., Hulliger, B., Kolb, J.-P., Lehtonen, R., Lussmann, D., Meraner, A., Myrskylä, M., Nedyalkova, D., Schoch, T., Templ, M., Valaste, M., Veijanen, A.: Policy Recommendations and Methodological Report. Research Project Report WP10 (D10.1/D10.2, FP7-SSH-2007-217322 AMELI). Available at: http://svn.uni-trier.de/AMELI (2011)
11. Särndal, C.-E.: The calibration approach in survey theory and practice. Survey Methodology, **33**, 99–119 (2007)
12. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer-Verlag, New York (1992)
13. Veijanen, A., Lehtonen, R.: Percentile-adjusted estimation of poverty indicators for domains under outlier contamination. Statistics in Transition **12**, 345–356 (2011)
14. Wu, C., Sitter, R.R.: A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association **96**, 185–193 (2001)