EXAMPLE Poverty rate estimation

Source: Lehtonen & Veijanen (2016a) Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) Analysis of Poverty Data by Small Area Estimation. Chichester: Wiley.

Assisting models

For poverty rate, the estimation of the proportions of poor people in the domains of interest can be interpreted as a problem of estimating the domain means of an indicator (binary) variable describing whether an individual is in poverty or not.

Therefore, model-assisted methods such as logistic GREG and model-assisted calibration are readily applicable.

Poverty indicator variable: $I\{y_k \le t\} = 1$ if person's equivalized income is smaller than or equal to the poverty threshold t (60% of the median equivalized income M in the population, according to EU definition), $I\{y_k \le t\} = 0$ otherwise

Estimation of median income

Distribution function of equivalized income variable y in the population:

$$F_U(t) = \frac{1}{N}\sum_{k\in U} I\{y_k \le t\}$$

HT type estimator of the population distribution function:

$$\hat{F}_U(t) = \frac{1}{\hat{N}}\sum_{k\in s} a_k I\{y_k \le t\}, \text{ where } \hat{N} = \sum\nolimits_{k\in s} a_k \text{ and } a_k = 1/\pi_k$$

Median equivalized income estimate $\hat{M}$ in the population: the smallest y_k ($k \in s$) for which $\hat{F}_U(y_k) > 0.5$

Assisting models for logistic GREG

(a) Fixed-effects model: Logistic fixed-effects regression model with domain-specific fixed intercepts for the binary poverty indicator $v_k = I\{y_k \le 0.6\hat{M}\}$:

$$p_k = P\{v_k = 1\} = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta})} \quad (k \in U) \tag{1}$$

where we include *domain-specific fixed intercepts* by including the domain membership indicators I_{dk} in the model:

$$\mathbf{x}_k = (I_{1k}, I_{2k}, ..., I_{Dk}, x_{1k}, x_{2k}, ..., x_{pk})'$$
$$\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, ..., \beta_{0D}, \beta_1, \beta_2, ..., \beta_p)'$$

(b) Mixed model: Logistic fixed-effects regression model with *domain-specific random intercepts* u_d following $N(0, \sigma_u^2)$:

$$p_k^{(m)} = P\{v_k = 1 \mid u_d\} = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)} \quad (k \in U_d), \tag{2}$$

where $\mathbf{x}_k = (1, x_{1k}, x_{2k}, ..., x_{pk})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$

Parameters $\boldsymbol{\beta}$ and σ_u^2 are first estimated from the sample using ML (R functions `glm` and `glmer`) and the random effects are then predicted for each domain to obtain the fitted values $\hat{p}_k$ and $\hat{p}_k^{(m)}$

GREG estimators

(a) GREG estimator assisted by logistic fixed-effects model (LGREG)

$$\hat{t}_{d;LGREG} = \sum_{k\in U_d} \hat{p}_k + \sum_{k\in s_d} a_k \left(v_k - \hat{p}_k\right) \tag{3}$$

Estimator for poverty rate:

$$\hat{r}_d = \frac{\hat{t}_{d;LGREG}}{N_d}$$

(b) GREG estimator assisted by logistic mixed model (MLGREG)

$$\hat{t}_{d;MLGREG} = \sum_{k\in U_d} \hat{p}_k^{(m)} + \sum_{k\in s_d} a_k \left(v_k - \hat{p}_k^{(m)}\right) \tag{4}$$

Estimator for poverty rate:

$$\hat{r}_d^{(m)} = \frac{\hat{t}_{d;MLGREG}}{N_d}$$

Simulation experiments

Data

Unit-level population of one million persons in 36 NUTS4 regions in Western Finland (register data from Statistics Finland)

In simulation experiments, $K = 1000$ SRSWOR samples of size $n = 1000$ were drawn

Study variable: Binary poverty indicator $v_k = I\left\{y_k \le 0.6\hat{M}\right\}$
Auxiliary variables: age class (5 classes), gender, labour force status and socio-economic status

The models included class indicators corresponding to main effects and interactions of age class with gender, labour force status and socio-economic status

Quality of an estimator $\hat{\theta}_d$ of θ_d ($d = 1,...,D$) over samples s_k ($k = 1,2,...,1000$) was assessed by absolute relative bias (ARB) and relative root mean squared error (RRMSE):

$$\text{ARB}\left(\hat{\theta}_d\right) = \frac{\left|\frac{1}{1000}\sum_{k=1}^{1000}\hat{\theta}_d\left(s_k\right) - \theta_d\right|}{\theta_d}; \quad \text{RRMSE}\left(\hat{\theta}_d\right) = \frac{\sqrt{\frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{\theta}_d\left(s_k\right) - \theta_d\right)^2}}{\theta_d}.$$

HT-CDF estimator of poverty rate is based on a HT type estimator of the distribution function in a domain:

$$\hat{F}_d(t) = \frac{1}{\hat{N}_d}\sum_{k\in s_d} a_k I\left\{y_k \le t\right\}$$

The poverty rate is then estimated by

$$\hat{r}_{d;HT} = \hat{F}_d\left(0.6\hat{M}\right) \tag{5}$$

We fitted two types of models:

(a) fixed effects logistic model with NUTS4 intercepts
(b) logistic mixed model with random intercepts associated with the NUTS4 domains

All estimators were nearly design unbiased as expected (Table 1).

Model choice had larger effect on RRMSE:

Fixed-effects logistic model with domain-specific intercepts did not yield good results with the model-assisted logistic GREG estimator LGREG. The reason might be instable estimation, in the group of smallest domains in particular (note: there are 36 fixed intercept parameters to be estimated). This result suggests that a fixed-effects model with domain-specific parameters might not be a good idea if the number of domains is large.

The best results were obtained with the logistic mixed model assisted MLGREG estimator. This estimator outperformed clearly the HT and LGREG estimators.

Table 1 Absolute relative bias (ARB) and relative root men squared error (RRMSE) of estimators of poverty rate in a design-based simulation experiment of 1,000 SRSWOR samples.

	Estimator	ARB (%)			RRMSE (%)		
		Expected domain sample size			Expected domain sample size		
		5-12	12-25	25-151	5-12	12-25	25-151
Direct estimator	HT (5)	1.7	2.2	0.9	83.7	60.1	38.9
Indirect estimators *Assisting models*							
(a) Fixed-effects logistic model (1) with domain-specific intercepts	LGREG (3)	1.8	1.9	0.9	83.7	59.9	38.5
(b) Mixed logistic model with domain-specific random intercepts (2)	MLGREG (4)	2.0	1.8	0.9	72.4	55.0	36.8