S.A.M.P.L.E.

Small Area Methods for Poverty and Living condition Estimates

Dissemination

WP2 Small area estimation of poverty and inequality indicators

Universita di Pisa
University of Manchester
Universidad Carlos III de Madrid
Universidad Miguel Hernández de Elche

SEVENTH FRAMEWORK PROGRAMME

EU – FP7 - SSH-2007-1
Grant Agreement no 217565

Índice

Description of WP2

The partners

1: Universita di Pisa
3: University of Manchester
4: Universidad Carlos III de Madrid
5: Universidad Miguel Hernández de Elche

The Tasks

2.1: Estimate the c.d.f of income at small area level (1, 3)
2.2: Small area estimates of poverty with spatial models (1, 3, 4)
2.3: SAE of poverty with temporal models (5)
2.4: SAE of poverty with spatio-temporal models (1, 4, 5)

Description of WP2

2.1: Estimation of c.d.f of income at small area level

- WP2 investigates method to estimate the Cumulative Distribution Function of Income (CDFI) in each unplanned domain (total disposable household income, equivalised total disposable income - EU definition of income and modified OECD scale).

- WP2 intends to utilize M-quantile models for small area estimation.

- WP2 performs the estimation of the cumulative distribution function of the variable of interest by combining both M-quantile and random effects models with appropriate model unbiased and design consistent estimators of the distribution function.

Description of WP2

2.1-2.4: Small area estimates of poverty indicators

2.1. WP2 proposes new methodologies for estimating poverty and inequality indicators along with their accuracy measures in small areas.

(2.1a) WP2 develops small area estimates of poverty indicators that take into account the spatial correlation between neighbour areas.

(2.1b) WP2 develops small area estimates of poverty indicators using M-quantile Geographically Weighted Regression model.

2.2. WP2 develops small area estimates using data from different periods through models that "borrow strength from time".

2.3. WP2 develops small area estimates through spatial-temporal models, which "borrow strength from space and time"

Poverty measures

- Let E_{dj} be a quantitative measure of welfare for unit j in area d.

- For example $E_{dj} = R_{dj}/H_{dj}$, where
$R_{dj} =$ total net monetary income of household j and area d,
$H_{dj} =$ total number of normalized members of household j and area d,

$$H_{dj} = 1 + 0.5(H_{dj \geq 14} - 1) + 0.3 H_{dj<14},$$

$H_{dj \geq 14}$ is the number of members aged 14 or more in (d, j),
$H_{dj<14}$ is the number of members aged 13 or less in (d, j).

- Let z be the poverty line; that is, the threshold for E_{dj} under which a person is considered as "under poverty".

- The family of poverty measures of Foster, Greer and Thorbecke (1984), called FGT poverty measures, for a small area d is

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

Poverty measures

- Note that

$\hookrightarrow I(E_{dj} < z) = 1$ if $E_{dj} < z$ (person under poverty)
$\hookrightarrow I(E_{dj} < z) = 0$ 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 measure for $\alpha = 1$ is called poverty gap, and measures the small area mean of the relative distance to non-poverty (the poverty gap) of each individual.

- For $\alpha = 2$ the measure is called poverty severity.

Direct estimators of poverty measures

- The direct estimators of the FGT measures are

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

where

$\hookrightarrow$ $\hat{N}_d = \sum_{j \in s_d} w_{dj}$ is the direct estimator of the population size N_d of small area d.

$\hookrightarrow$ w_{dj} is the sampling weight (inverse of the probability of inclusion) of individual j in the sample from small area d

EB method for poverty estimation

- **Assumption:** There exists a transformation $Y_{dj} = T(E_{dj})$ of the welfare variables E_{dj} which follows a normal distribution.

- Poverty measure as a function of transformed variables:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left\{ \frac{z - T^{-1}(Y_{dj})}{z} \right\}^{\alpha} I\left\{ T^{-1}(Y_{dj}) < z \right\} = h_{\alpha}(\boldsymbol{y}_d),$$

where $\boldsymbol{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$.

- **Best estimator:** The estimator of $F_{\alpha d}$ with minimum MSE is

$$\hat{F}_{\alpha d}^{B} = E_{\boldsymbol{y}_{dr}}\left[F_{\alpha d} | \boldsymbol{y}_{ds}\right], \quad F_{\alpha d} = h_{\alpha}(\boldsymbol{y}_d),$$

where $\boldsymbol{y}_{ds}$ and $\boldsymbol{y}_{dr}$ denote respectively sample and out-of-sample parts of $\boldsymbol{y}_d$.

EB method for poverty estimation

- **Empirical Best (EB) estimator:** Expectation calculated with respect to the distribution of $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$ with estimated unknown parameters.

- **Nested error linear model:**

$$Y_{dj} = \mathbf{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \ d = 1, \ldots, D.$$
$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2).$$

- Distribution of $\boldsymbol{y}_d$:

$$\boldsymbol{y}_d \overset{ind}{\sim} N(\boldsymbol{\mu}_d, \mathbf{V}_d), \quad d = 1 \ldots, D,$$

where

$$\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta} \text{ and } \mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \sigma_e^2 I_{N_d}.$$

EB method for poverty estimation

- Decomposition in sample and out-of-sample:

$$\boldsymbol{\mu}_d = \begin{pmatrix} \boldsymbol{\mu}_{ds} \\ \boldsymbol{\mu}_{dr} \end{pmatrix}, \quad \mathbf{V}_d = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}$$

- Distribution of $\boldsymbol{y}_{dr}$ given $\boldsymbol{y}_{ds}$:

$$\boldsymbol{y}_{dr} | \boldsymbol{y}_{ds} \sim N(\boldsymbol{\mu}_{dr|ds}, \mathbf{V}_{dr|ds}),$$

where

$$\begin{aligned} \boldsymbol{\mu}_{dr|ds} &= \boldsymbol{\mu}_{dr} + \mathbf{V}_{drs}\mathbf{V}_{ds}^{-1}(\boldsymbol{y}_{ds} - \boldsymbol{\mu}_{ds}), \\ \mathbf{V}_{dr|ds} &= \mathbf{V}_{dr} - \mathbf{V}_{drs}\mathbf{V}_{ds}^{-1}\mathbf{V}_{dsr}. \end{aligned}$$

EB method for poverty estimation

- **Monte Carlo approximation of best estimator:**

(a) Generate L non-sample vectors $\boldsymbol{y}_{dr}^{(\ell)}$, $\ell = 1, \dots, L$ from the conditional distribution of $\boldsymbol{y}_{dr} | \boldsymbol{y}_{ds}$.

(b) Attach the sample elements to form a population vector $\boldsymbol{y}_d^{(\ell)} = (\boldsymbol{y}_{ds}, \boldsymbol{y}_{dr}^{(\ell)})$, $\ell = 1, \dots, L$.

(c) Calculate the poverty measure with each population vector $F_{\alpha d}^{(\ell)} = h_\alpha(\boldsymbol{y}_d^{(\ell)})$, $\ell = 1, \dots, L$. Then take the average over the L Monte Carlo generations:

$$\hat{F}_{\alpha d}^{B} = E_{\boldsymbol{y}_{dr}}[F_{\alpha d} | \boldsymbol{y}_{ds}] \cong \frac{1}{L} \sum_{\ell=1}^{L} F_{\alpha d}^{(\ell)}.$$

Time mixed models

- Unit-level linear mixed models

$$y_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,dt} + w_{dtj}^{-1/2} e_{dtj}, \qquad \begin{array}{l} d = 1, \ldots, D, \\ t = 1, \ldots, m_d, \\ j = 1, \ldots, n_{dt}. \end{array} \tag{1}$$

where

(TM1) $u_{1,d}$ i.i.d. $N(0, \sigma_1^2)$, $(u_{2,d1}, \ldots, u_{2,dm_d})$ i.i.d. $AR(1; \sigma_2^2, \rho)$ and e_{dtj} i.i.d. $N(0, \sigma_0^2)$ are independent.

(TM2) $u_{1,d}$ i.i.d. $N(0, \sigma_1^2)$, $u_{2,dt}$ i.i.d. $N(0, \sigma_2^2)$ and e_{dtj} i.i.d. $N(0, \sigma_0^2)$ are independent.

Time mixed models

- Area-level linear mixed models

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, m_d,$$

where
$\hookrightarrow$ y_{dt} is a direct estimator of the characteristic of interest and
$\hookrightarrow$ $\mathbf{x}_{dt}$ is a vector containing the population (aggregated) values of p auxiliary variables.

(TM3) $(u_{d1}, \ldots, u_{2dm_d})$ i.i.d. $AR(1; \sigma_u^2, \rho)$ and $e_{dt} \stackrel{ind}{\sim} N(0, \sigma_{dt}^2)$ are independent.

(TM4) u_{dt} i.i.d. $N(0, \sigma_u^2)$ and $e_{dt} \stackrel{ind}{\sim} N(0, \sigma_{dt}^2)$ are independent.

Spatial mixed models

- Unit-level mixed models

$$y_{dj} = \mathbf{x}_{dj}\boldsymbol{\beta} + v_d + w_{dj}^{-1/2} e_{dj}, \quad d = 1, \ldots, D, \; j = 1, \ldots, n_d \tag{2}$$

where

(SM1) $(v_1, \ldots, v_D)$ i.i.d. $SAR(1; \sigma_v^2, \rho, \mathbf{P})$ and e_{dj} i.i.d. $N(0, \sigma_e^2)$ are independent.

- Area-level linear mixed models

$$y_d = \mathbf{x}_d\boldsymbol{\beta} + v_d + e_d, \quad d = 1, \ldots, D,$$

where

$\hookrightarrow$ y_d is a direct estimator of the characteristic of interest and

$\hookrightarrow$ $\mathbf{x}_d$ is a vector containing the population (aggregated) values of p auxiliary variables.

(SM2) $(v_1, \ldots, v_D)$ i.i.d. $SAR(1; \sigma_v^2, \rho, \mathbf{P})$ and $e_d \overset{ind}{\sim} N(0, \sigma_d^2)$ are independent.

spatio-temporal mixed models

- Area-level spatio-temporal linear mixed models

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{1d} + u_{2dt} + e_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T,$$

where

$\hookrightarrow$ y_{dt} is a direct estimator of the characteristic of interest and

$\hookrightarrow$ $\mathbf{x}_{dt}$ is a vector containing the population (aggregated) values of p auxiliary variables.

(STM1) $\{u_{1d}\}$, $\{u_{2dt}\}$ and $\{e_{dt}\}$ are independent with distributions $\{u_{1d}\}_{d=1}^{D} \sim SAR(1)$, $\{u_{2dt}\}$ i.i.d $N(0, \sigma_2^2)$ and $e_{dt} \sim N(0, \sigma_{dt}^2)$.

(STM2) $\{u_{1d}\}$, $\{u_{2dt}\}$ y $\{e_{dt}\}$ are independent with distributions $\{u_{1d}\}_{d=1}^{D} \sim SAR(1)$, $\{u_{2dt}\}_{t=1}^{T}$ i.i.d $AR(1)$ and $e_{dt} \sim N(0, \sigma_{dt}^2)$.

M-quantile models

- With regression models we model the mean of the variable of interest (y) given the covariates (x)

- A more complete picture is offered by modeling not only the mean of y given x, but also the quantiles.

- Examples include the median, the 25th, 75th percentiles.

- This is known as quantile regression

- An M-quantile regression model for quantile q is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}(q) + \mathrm{e}(q)$$

where q is a-priori chosen.

M-quantile models

- Estimate of $\boldsymbol{\beta}(q)$ is obtained via Iterative Weighted Least Squares:

$$\widehat{\boldsymbol{\beta}}(q) = (\boldsymbol{X}^t\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\mathbf{W}\boldsymbol{y}$$

- $\mathbf{W}$ is an $n \times n$ diagonal weighting matrix that depends on both the influence function and the quantile we are modeling

- Central Idea: Area effects can be described by estimating an area specific q value ($\widehat{\theta}_d$) for each area (group) of a hierarchical dataset

- Estimate the area specific target parameter by fitting an M-quantile model for each area at $\widehat{\theta}_d$

$$y_{dj} = \mathbf{x}_{dj}\widehat{\boldsymbol{\beta}}(\widehat{\theta}_d) + e_{dj}(\widehat{\theta}_d)$$

- A mixed model uses random effects u_d to capture the dissimilarity between groups. M-quantile models attempt to capture this dissimilarity via the group-specific M-quantile coefficients $\widehat{\theta}_d$

Estimation of cumulative distribution functions

- Estimation of the distribution function of income will be performed using both M-quantile and random effects models (see Chambers and Dunstan 1986 ; Rao, Kovar and Mantel 1990).

- The CDF estimator can be further used for estimating other quantiles of the small area distribution function of the variable of interest.

- This is achieved by integrating the CDF estimator

$$\int_{-\infty}^{q} t \, d\widehat{F}_d(t)$$

References

Chambers, R. and Dunstan, R. (1986). Estimating distribution function from survey data, Biometrika, **73**, 597-604.

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

Pfeffermann, D., Terryn, B. and Moura, F. (2007). *Bayesian small area estimation of literacy under a two part random effects model*, Presentation at SAE2007 Conference held in Pisa.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, **2**, 365-375.