On influence of clustering population on estimation accuracy of population totals vector

Janusz L. Wywiał, Grzegorz Sitek

Department of Statistics, Econometrics and Mathematics
University of Economics in Katowice, Poland,
janusz.wywial@ue.katowice.pl, grzegorz.sitek@ue.katowice.pl

ITACOSM 2019, 05-07 June 2019

Contents:

Simple cluster sample vector estimator
Basic notation

- U is population of size N,
- m is number of variables observed in$U$,
- $\boldsymbol{y}_k = [y_{k,1} ... y_{k,m}]$ where $k \in U$,
- $\mathcal{D} = \{U_1, ..., U_h, ..., U_G\}$ is partition of U into clusters U_h, $h = 1, .., G$, $\bar{N} = N/G$,

$$\bar{\boldsymbol{y}} = [\bar{y}_1 ... \bar{y}_m] = \sum_{k \in U} \boldsymbol{y}_k / N, \quad \boldsymbol{y}_U = N\bar{\boldsymbol{y}} = \sum_{k \in U} \boldsymbol{y}_k = [y_{U,1} ... y_{U,m}],$$

$$y_{U,i} = \sum_{k \in U} y_{k,i}, \quad \boldsymbol{C} = [c_{i,j}], \quad c_{i,j} = \sum_{k \in U} (y_{k,i} - \bar{y}_i)(y_{k,j} - \bar{y}_j)/(N-1),$$

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2} \boldsymbol{C} \boldsymbol{D}^{-1/2} = [r_{i,j}], \quad \boldsymbol{D} = [v_i], \quad r_{i,j} = \frac{c_{i,j}}{\sqrt{v_i v_j}}, \quad v_i = c_{i,i}.$$

Simple cluster sample vector estimator

Basic notation

$$\bar{\boldsymbol{y}}_{U_h} = \sum_{k \in U_h} \boldsymbol{y}_k / N_h, \quad \bar{\boldsymbol{y}}_{U_h} = [\bar{y}_{U_h,1} \ldots \bar{y}_{U_h,m}], \quad \bar{y}_{U_h,i} = \sum_{k \in U_h} y_{k,i} / N_h,$$

$$\boldsymbol{y}_{U_h} = N_h \bar{\boldsymbol{y}}_{U_h} = \sum_{k \in U_h} \boldsymbol{y}_k, \quad \boldsymbol{y}_{U_h} = [y_{U_h,1} \ldots y_{U_h,m}], \quad y_{U_h,i} = \sum_{k \in U_h} y_{k,i},$$

$$\boldsymbol{C}_{U_h} = [c_{U_h,i,j}], \quad c_{U_h,i,j} = \sum_{k \in U_h} (y_{k,i} - \bar{y}_{U_h,i})(y_{k,j} - \bar{y}_{U_h,j}) / (N_h - 1),$$

$$\bar{\boldsymbol{y}}_{\mathcal{D}} = \sum_{h=1}^{G} \boldsymbol{y}_{U_h} / G = \boldsymbol{y}_U / G = [\bar{y}_{\mathcal{D},1} \ldots \bar{y}_{\mathcal{D},m}], \quad \bar{y}_{\mathcal{D},i} = \sum_{h=1}^{G} y_{U_h,i} / G = y_{U,i} / G,$$

$$\boldsymbol{y}_{\mathcal{D}} = G \bar{\boldsymbol{y}}_{\mathcal{D}} = \sum_{h=1}^{G} = \boldsymbol{y}_{U_h} = \boldsymbol{y}_U, \quad \boldsymbol{C}_{\mathcal{D}} = [c_{\mathcal{D},i,j}],$$

$$c_{\mathcal{D},i,j} = \sum^{G} (y_{U_h,i} - \bar{y}_{\mathcal{D},i})(y_{U_h,j} - \bar{y}_{\mathcal{D}}, j) / (G - 1), \quad i, j = 1, \ldots, m.$$

Simple cluster sample vector estimator

Properties of the stimator

$$\tilde{\boldsymbol{y}}_S = \frac{G}{g}\sum_{h\in s}\sum_{k\in U_h}\boldsymbol{y}_k = \frac{G}{g}\sum_{h\in s}\boldsymbol{y}_{U_h}, \quad \boldsymbol{V}(\tilde{\boldsymbol{y}}_S) = \frac{G(G-g)}{g}\boldsymbol{C}_{\mathcal{D}}, \tag{1}$$

$$\boldsymbol{V}(\tilde{\boldsymbol{y}}_S) = \frac{G(G-g)}{g}\bar{N}\boldsymbol{C}\left(\boldsymbol{I}_m + \frac{N-G}{G-1}\boldsymbol{\Delta}\right) + \frac{G(G-g)}{g}\boldsymbol{A} \tag{2}$$

$$\text{homogeneity matrix:} \qquad \boldsymbol{\Delta} = \boldsymbol{I}_m - \boldsymbol{C}^{-1}\boldsymbol{C}_*, \tag{3}$$

$$\boldsymbol{A} = [a_{i,j}]; \quad a_{i,j} = \frac{1}{G-1}\sum_{h=1}^{G}(N_h - \bar{N})N_h\bar{y}_{U_h,i}\bar{y}_{U_h,j}, \tag{4}$$

$$\boldsymbol{C}_* = [c_{*i,j}], \quad c_{*i,j} = \frac{1}{N-G}\sum_{h=1}^{G}\sum_{k\in U_h}(y_{k,i}-\bar{y}_{U_h,i})(y_{k,j}-\bar{y}_{U_h,j}), \tag{5}$$

The eigenvalues of $\boldsymbol{\Delta}$ and the diagonal elements of $\boldsymbol{\Delta}$ take values from $\left[-\frac{G-1}{N-G}; 1\right]$.

Simple cluster sample vector estimator

Relative efficiency

The relative efficiency coefficient (Rao and Scott (1981):

$$deff(\tilde{\boldsymbol{y}}_S) = \lambda\left(\boldsymbol{V}(\boldsymbol{y}_S)^{-1}\boldsymbol{V}(\tilde{\boldsymbol{y}}_S)\right) \propto \lambda\left(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{D}}\right). \tag{6}$$

where $\lambda(...)$ is maximal eigenvalue and ordinary estimator

$$\boldsymbol{y}_S = \frac{N}{n}\sum_{k\in S}\boldsymbol{y}_k, \qquad \boldsymbol{V}(\boldsymbol{y}_S) = \frac{N(N-n)}{n}\boldsymbol{C} \tag{7}$$

Estimator $\tilde{\boldsymbol{y}}_S$ is not worst than $\boldsymbol{y}_S$ if and only if $\boldsymbol{V}(\tilde{\boldsymbol{y}}_S) - \boldsymbol{V}(\boldsymbol{y}_S)$ is non-positive definite and all eigenvalues of $\boldsymbol{V}(\boldsymbol{y}_S)^{-1}\boldsymbol{V}(\tilde{\boldsymbol{y}}_S)$ take values from [0; 1].

$$deff(\tilde{\boldsymbol{y}}_S) = 1 + \lambda\left(\frac{N-G}{G-1}\boldsymbol{\Delta} + \frac{1}{\bar{N}}\boldsymbol{C}^{-1}\boldsymbol{A}\right). \tag{8}$$

If $N_h = const$ for all $h = 1, ..., G$:

$$0 \leq deff(\tilde{\boldsymbol{y}}_S) = 1 + \frac{N-G}{G-1}\lambda(\boldsymbol{\Delta}) \leq \frac{N-1}{G-1}. \tag{9}$$

- Let $\boldsymbol{y}_k > \boldsymbol{0}$ for all $k = 1, ..., N$,
- evaluation of squared distances $d_k = \boldsymbol{y}_k \boldsymbol{y}_k^T$ of $\boldsymbol{y}_k$ from the zero vector $\boldsymbol{0}$ for all $k \in U$,
- let us assume that $d_k \leq d_{k+1}$ for $k = 1, ..., N-1$,
- h-th cluster is identified by such $k \in U_h$ that $k = (i-1)G + h$, for $i = 1, ..., M$ and $h = 1, ..., G$,
- this leads to: $d_{U_h} \leq d_{U_{h+1}}$ for $h = 1, ..., G-1$ where $d_{U_h} = \sum_{k \in U_h} d_k$.

- Let $d_k = (\boldsymbol{y}_k - \bar{\boldsymbol{y}})(\boldsymbol{y}_k - \bar{\boldsymbol{y}})^T$ be the squared distance of $\boldsymbol{y}_k$ from vector $\bar{\boldsymbol{y}}$ for all $k \in U$,
- let $d_k \leq d_{k+1}$ for $k = 1, ..., N-1$.
- when M is even and $N = MG$, then

$$U_h = \{(h-1)\frac{M}{2} + i; N - (h-1)\frac{M}{2} - i + 1\}$$

for $h = 1, ..., G$ and $i = 1, ...M/2$.

- Particularly, if $M = 2$ and $N = MG$,

$$U_h = \{h; N - h + 1\}$$

for $h = 1, ..., G$.

- Let $\mathcal{D}^{(0)} = \{U_1^{(0)}, ..., U_G^{(0)}\}$ be any start partition of population into clusters of the same sizes,
- in the t-h (t=0,1,...) iteration partition $\mathcal{D}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$ is generated through permutation population elements of U,
- for assumed $t = T$, $\mathcal{D}_3$ is treated as optimal when

$$\mathcal{D}_3 = arg(min_{\{t=1,..,T\}}(\lambda(\mathbf{\Delta}(\mathcal{D}^{(t)})))). \qquad (10)$$

Clustering algorithms

Algorithm $\mathcal{D}_4$

- Let $\mathcal{D}^{(0)} = \{U_1^{(0)}, ..., U_G^{(0)}\}$ be any start partition of the population into clusters of not necessary of the same size,
- let $f : U \to \mathcal{D}^{(t)}$, $f_t(k) = h$, if and only if $k \in U_h^{(t)}$.
- in iteration $t + 1$ we randomly choice number k_* from $1, ..., N$,
- element k_* is moved from the cluster $h_\# = f_t(k_*)$ to cluster h_*. h_* is randomly drawn from $\{h : h = 1, ..., G; h \neq h_\#\}$. This leads to new partition $\mathcal{D}^{(t+1)}$,
- let $\lambda_{t+1} = \lambda(\boldsymbol{C}_{\mathcal{D}^{(t+1)}})$. If $\lambda_{t+1} < \lambda_t$, then $\mathcal{D}^{(t+1)}$ is the current partition and we start the iteration $t + 2$ of the algorithm,
- if $\lambda_{t+1} \geq \lambda_t$, then we start the stage $t + 2$ of the algorithm from partition $\mathcal{D}^{(t)}$;
- the algorithm is stopped when number of the iteration reach assumed level T;
- this algorithm minimizes $deff(\tilde{\boldsymbol{y}}_S)$.

- $\mathcal{D}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$ is the resulted of t-th iteration where $t = (l-1)N + k$, $k = 1, ..., N$, $l = 1, 2, ...$;
- let $\lambda_t = \lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}})$ and let $f : U \to \mathcal{D}^{(t)}$, $f_t(l) = h \Leftrightarrow l \in U_h^{(t)}$;
- in stage $t+1$ element $k \in U_h^{(t)}$, where $h = f_t(k)$, is moved to clusters $U_z^{(t)}$, $z \neq h$, $z = 1, ..., G$ and calculated the following

$$(k, \underline{z}) = arg\left(min_{\{z=1,...,G, z \neq f_t(k)\}}\left(\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(k, z))\right)\right) \qquad (11)$$

- $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(k, z))$ is evaluated for the partition $\mathcal{D}^{(t)}$ in which clusters $U_z^{(t)}$ and $U_h^{(t)}$ are replaced by $\{U_z^{(t)} \cup \{k\}\}$ and $\{U_h^{(t)} - \{k\}\}$, respectively, and $h = f_t(k)$;

- If $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(\underline{z})) < \lambda_t$, then $\lambda_{t+1} = \lambda(\boldsymbol{C}_{\mathcal{D}^{(t+1)}})$ and $\mathcal{D}^{(t+1)}$ is equal to $\mathcal{D}^{(t)}$ where clusters $U_{\underline{z}}^{(t)}$ and $U_{\underline{h}}^{(t)}$ are replaced by $U_{\underline{z}}^{(t+1)} = \{U_{\underline{z}}^{(t)} \cup \{k\}\}$ and $U_h^{(t+1)} = \{U_h^{(t)} - \{k\}\}$, respectively;
- if $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(\underline{z})) \geq \lambda_t$, then $\mathcal{D}^{(t+1)} = \mathcal{D}^{(t)}$ and $\lambda_{t+1} = \lambda_t$;
- the iteration process is stopped when $\lambda_{t+N} = \lambda_t$ or the number of the iterations attains the preassigned level T.

- in iteration $t+1$ element $k \in U_h^{(t)}$, where $h = f_t(k)$, is moved to clusters $U_z^{(t)}$, $z \neq h$, $z = 1, ..., G$ and calculated:

$$(\underline{k}, \underline{z}) = arg\left(min_{\{k \in U\}} min_{\{z \neq f_t(k), z=1,...,G\}}\left(\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(k,z))\right)\right) \tag{12}$$

- $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(k,z))$ is evaluated for $\mathcal{D}^{(t)}$ in which clusters $U_z^{(t)}$ and $U_h^{(t)}$ are replaced by $\{U_z^{(t)} \cup \{k\}\}$ and $\{U_h^{(t)} - \{k\}\}$, respectively, and $h = f_t(k)$;
- if $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(\underline{k},\underline{z})) < \lambda_t$, then $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t+1)}}) = \lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(\underline{k},\underline{z}))$ and $\mathcal{D}^{(t+1)}$ is equal to $\mathcal{D}^{(t)}$ where $U_z^{(t)}$ and $U_h^{(t)}$ are replaced by $U_{\underline{z}}^{(t+1)} = \{U_{\underline{z}}^{(t)} \cup \{k\}\}$ and $U_h^{(t+1)} = \{U_h^{(t)} - \{k\}\}$, respectively;
- the process is stopped when $\lambda(\boldsymbol{C}_{\mathcal{D}^{(t)}}(\underline{k},\underline{z})) \geq \lambda_t$.

Accuracy analysis

Data on Swedish municipalities from Särndal C.E., et al. Variables y_1 and y_2 are the real estate values and number of municipal employees, respectively.

Table 1. Relative efficiencies.

n	(M,g)	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$
1	2	3	5	6	7	8	9
16	(2,8)	0.99	1.11	0.82	0.56	0.64	0.77
16	(4,4)	1.10	2.15	0.75	0.18	0.44	0.58
16	(8,2)	1.17	4.25	0.62	0.05	0.18	0.44
28	(2,14)	0.99	1.11	0.82	0.56	0.64	0.77
28	(4,7)	1.10	2.15	0.75	0.18	0.44	0.58
28	(14,2)	1.31	7.35	0.50	0.02	0.04	0.20
48	(2,24)	0.99	1.11	0.82	0.56	0.64	0.77
48	(4,12)	1.10	2.15	0.75	0.18	0.44	0.58
48	(8,6)	1.17	4.25	0.62	0.05	0.18	0.44

Source: Own calculations.

Conclusions

- Only under $\mathcal{D}_1$ and $\mathcal{D}_2$ the accuracy of $\boldsymbol{y}_S$ is not less than the accuracy $\tilde{\boldsymbol{y}}_S$ for all (M, g).
- $\mathcal{D}_4$ leads to the most efficient estimation based on $\tilde{\boldsymbol{y}}_S$.
- $\mathcal{D}_3$ leads to the most efficient estimation based on $\tilde{\boldsymbol{y}}_S$, when we assume that the population is split into clusters of the same sizes.
- For $\mathcal{D}_1$ and $\mathcal{D}_2$ the efficiency of $\tilde{\boldsymbol{y}}_S$ decreases, when number of clusters g decreases under fixed n.
- For $\mathcal{D}_3$-$\mathcal{D}_6$ the efficiency of $\tilde{\boldsymbol{y}}_S$ increases, when number of clusters g decreases under fixed n.
- For instance, under $\mathcal{D}_4$, when $(M, g) = (2, 14)$ and $(M, g) = (14, 2)$, $deff(\tilde{\boldsymbol{y}}_S) = 0.56$ and $deff(\tilde{\boldsymbol{y}}_S) = 0.02$, respectively.

Reference

- Borovkov A.A. (1981). Mathematical Statistics. Estimation. Testing Hypotheses. (in Russian) Nauka Moskva.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons, Inc. New York.
- Rao, J.N.K., Scott A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests of goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* vol. 76, no. 374, pp. 221-230.
- Särndal C.E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling* Springer-Verlag, New York.
- Wywiał J.L. (2003). *Some Contributions to Multivariate Methods in Survey Sampling.* University of Economics in Katowice. $http://www.ue.katowice.pl/fileadmin/user\_upload/wydawnictwo/Darmowe\_E-Booki$

Thank you very much