A Multivariate Regression Estimator for Rotating Sampling Surveys

Karen Caruana (kc12g13@soton.ac.uk)[1] and Yves G. Berger (y.g.berger@soton.ac.uk)[1]

Keywords: Design-based approach, multivariate regression, regression estimator, calibration, correlation

1. INTRODUCTION

Longitudinal surveys collect information on several occasions, or time points [1], [2]. Consider that we have two occasions or waves labelled 1 and 2. The samples selected on occasions 1 and 2 are rarely completely overlapping samples, as not all the units are selected on both occasions. It is common practice to have a large fraction of units sampled at both occasions. Surveys which have this feature are called rotating sampling surveys.

The customary point estimators are the Horvitz Thompson [3] and generalised regression [4] estimators of a total or a mean. We propose a new regression estimator for cross-sectional totals and change between totals. This estimator uses the information from both occasions simultaneously instead of each occasion separately. This estimator incorporates the auxiliary variables similar to the general regression estimator and the sample design variables specifying the rotating sampling design. The proposed estimator is multivariate because it combines the auxiliary information from the first and second occasion.

Longitudinal surveys are used to monitor change between population target parameters. For social policy makers, the estimation of change over time of social indicators as such youth employment rate, literacy rate and social deprivation indicators may be as important as cross-sectional indicators. The variance of change, for rotating sampling surveys, is a challenging subject since it requires to estimate correlations. Several authors proposed different estimators for correlations [5], [6], [7], [8] and [9]. A variance of change is proposed by extending the estimator proposed by [9] where besides the design variables, the auxiliary variables are included.

In the simulation study, the proposed estimator is compared with the Horvitz Thompson (HT) and generalised regression estimators. The relative bias and ratio of relative mean square errors are computed for the estimator of totals. We consider different correlations between the response variables and the auxiliary variables.

2. METHODS

In rotating sampling designs, a fixed proportion of sample units are replaced by new units at each wave. Each unit remains in the sample for the same number of waves [2].

Let s_1 and s_2 be the probability samples for the first occasion (selected from population U_1) and for the second occasion (selected from population U_2) respectively. Let s_{12} be

[1] University of Southampton, UK

the sample of units that are both in s_1 and s_2. Suppose that the sample size is fixed for both occasions. We consider that s_1 is composed of n_1 units with first-order inclusion unequal probabilities $\pi_{1;i} = pr\{i \in s_1\}$, where $pr\{.\}$ denotes the probability with respect to the design. Similarly, s_2 is composed of n_2 units. The n_2 units are selected with conditional inclusion unequal probabilities $\pi_{2;i}(s_1) = pr\{i \in s_2 | s_1\}$ which are such that n_c units are contained in s_c; where $s_c = s_1 \cap s_2$. Thus, the second wave inclusion probabilities are given by $\pi_{2;i} = E_1[\pi_{2;i}(s_1)]$; where $E_1[.]$ denotes the design expectation with respect to the first wave design. Finally, assume that for both waves, the sampling fractions are negligible; that is, $1 - \pi_{l;i} \approx 1$.

Let $= (\boldsymbol{y}_1, \boldsymbol{y}_2)^T$; where $\boldsymbol{y}_l = \left(y_{l;1}, y_{l;2}, \cdots, y_{l;n_l}\right)^T$, be the responses for the variable of interest for wave $l = 1, 2$. Define $\breve{\boldsymbol{Y}} = (\breve{\boldsymbol{y}}_1, \breve{\boldsymbol{y}}_2)^T$ where $\breve{\boldsymbol{y}}_l = \left(y_{l;1}\,\pi_{l;i}^{-1}, y_{l;2}\pi_{l;2}^{-1}, \cdots, y_{l;n_l}\pi_{l;nl}^{-1}\right)^T$. Let $\widehat{\boldsymbol{y}}$ be the vector of the HT estimators of the response variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$:

$$\widehat{\boldsymbol{y}} = (\hat{y}_1, \quad \hat{y}_2)^T,$$

where $\hat{y}_l = \sum_{i=1}^{n_l} y_{l;i}\pi_{l;i}^{-1}$.

Assume that J auxiliary variables are available for both waves. The vector of auxiliary variables of the k^{th} element of wave l is defined as: $\boldsymbol{x}_{l;k} = \left(x_{l;1;k}, x_{l;2;k}, \cdots, x_{l;J;k}\right)^T$. Let $\breve{\boldsymbol{X}}_l = \left[\breve{\boldsymbol{x}}_{l;1}, \breve{\boldsymbol{x}}_{l;2}, \cdots, \breve{\boldsymbol{x}}_{l;n}\right]$ where $\breve{\boldsymbol{x}}_{l;i} = \left(x_{l;1;k}\pi_{l;1;k}^{-1}, x_{l;2;k}\pi_{l;2;k}^{-1}, \cdots, x_{l;J;k}\pi_{l;J;k}^{-1}\right)^T$. Let $\boldsymbol{x} = (\boldsymbol{x}_1^T, \quad \boldsymbol{x}_2^T)^T$ be the $(2J$ x 1) vector of population totals of the auxiliary variables, where $\boldsymbol{x}_l = \left(\sum_{i\in U_l} x_{l;1;i}, \sum_{i\in U_l} x_{l;2;i}, \cdots, \sum_{i\in U_l} x_{l;J;i}\right)^T$ and the corresponding HT estimator vector is $\widehat{\boldsymbol{x}} = (\widehat{\boldsymbol{x}}_1^T, \quad \widehat{\boldsymbol{x}}_2^T)^T$; $\widehat{\boldsymbol{x}}_l = \left(\sum_{i\in s_l} x_{l;1;i}\pi_{l;1;i}^{-1}, \sum_{i\in s_l} x_{l;2;i}\pi_{l;2;i}^{-1}, \cdots, \sum_{i\in s_l} x_{l;J;i}\pi_{l;J;i}^{-1}\right)^T$.

Let the design variables be $z_{1;i} = \delta\{i \in s_1\}$ and $z_{2;i} = \delta\{i \in s_2\}$, where $\delta\{A\}$ is one when A is true and zero otherwise. Let define the matrix of the design variables as

$$\boldsymbol{Z}_s = (\boldsymbol{z}_1, \quad \boldsymbol{z}_2, \quad \boldsymbol{z}_c)^T$$

where $(z_{l;1}, \quad z_{l;2} \quad, \cdots, \quad z_{l;n_{12}})^T$ and $\boldsymbol{z}_c = (z_{1;1}z_{2;1}, \quad z_{1;2}z_{2;2}, \quad \cdots \quad, z_{1;n_{12}}z_{2;n_{12}})^T$; $n_{12} = n_1 + n_2 - n_c$. Define $\widehat{\boldsymbol{\gamma}}_s = (\widehat{\boldsymbol{x}}^T, \quad \widehat{\boldsymbol{z}}_s^T)^T$ and $\boldsymbol{\gamma}_U = (\boldsymbol{x}^T, \quad \boldsymbol{z}_U^T)^T$ as two $(2J +$ 3 x 1) vectors; where $\hat{\boldsymbol{z}}_s = \left(\sum_{i\in s_1} z_{1;i}, \sum_{i\in s_2} z_{2;i}, \sum_{i\in s_c} z_{c;i}\right)^T = (n_1, n_2, n_c)^T = \boldsymbol{z}_U$

The proposed multivariate generalised regression estimator is:

$$\widehat{\boldsymbol{y}}_s^{(PROP)} = \widehat{\boldsymbol{y}} + (\boldsymbol{\gamma}_U - \widehat{\boldsymbol{\gamma}}_s)^T\widehat{\boldsymbol{\beta}}_{XZ},$$

where $\widehat{\boldsymbol{\beta}}_{XZ} = \left(\breve{\boldsymbol{C}}_s^T\breve{\boldsymbol{C}}_s\right)^{-1}\left(\breve{\boldsymbol{C}}_s\right)^T\breve{\boldsymbol{Y}}$; $\breve{\boldsymbol{C}}_s = (\breve{\boldsymbol{X}}, \quad \boldsymbol{Z}_s)$, $\breve{\boldsymbol{X}} = (\breve{\boldsymbol{X}}_1, \quad \breve{\boldsymbol{X}}_2)$.

The multivariate regression estimator of change $\Delta = \sum_{i\in U_1} y_{1;i} - \sum_{i\in U_2} y_{2;i}$ is given by

$$\hat{\Delta} = (1, -1)\widehat{\boldsymbol{y}}_s^{(PROP)}.$$

The proposed variance of change is based upon [9] where the design variables and the auxiliary variables are included.

3. SIMULATION RESULTS

For the simulation study, we consider a population of N = 20,000 units. The sample size is the same for both waves, $n_1 = n_2 = 200$ and the number of sampling units common for both waves is n_c=120. The population is generated from a multivariate (i) normal distribution and (ii) lognormal distribution. 1,000 samples are selected using a random systematic sampling where the probabilities are unequal without replacement. We consider several correlations between the response variables and the auxiliary variables.

The proposed (PROP) estimator is compared with the HT and generalised regression (GREG) estimators. The relative bias (RB) and ratio of relative mean square error (RRMSE) are computed for the point estimators, cross-sectional variance and the variance of change. Tables 1 and 2 below shows the results for the data generated from the two different distributions considered.

The RB and RRMSE of the proposed point estimator is always smaller than the HT and GREG estimator. The RRMSE of the variance estimators are of a comparable order for normal distributions (see Table 1). With a log-normal distribution (see Table 2), the standard GREG estimator has the smallest RRMSE for the variance. We observe a small RB for the variance estimator for change of the proposed estimator.

Table 1: Results from data generated from a multivariate normal distribution

		RB			RRMSE		
Correlation		HT	GREG	PROP	HT	GREG	PROP
$\delta_{YY} = 0.2$,	$\hat{y}_1$	0.03	-0.01	-0.01	1.77	1.59	1.59
$\delta_{YX} = 0.2$,	$\hat{y}_2$	0.09	0.05	0.05	1.70	1.52	1.51
$\delta_{XX} = 0.2$.	$\widehat{var}(\hat{y}_1)$	-2.61	3.27	-3.47	11.08	11.02	10.69
	$\widehat{var}(\hat{y}_2)$	3.61	5.60	3.61	11.40	12.23	11.64
	$\widehat{var}(\hat{\Delta})$	11.55	0.99	-2.21	15.56	9.05	9.21
$\delta_{YY} = 0.8$,	$\hat{y}_1$	0.09	0.02	0.00	1.77	1.07	1.00
$\delta_{YX} = 0.8$,	$\hat{y}_2$	0.09	0.05	0.04	1.73	1.17	1.01
$\delta_{YX} = 0.8$.	$\widehat{var}(\hat{y}_1)$	-3.28	-1.14	-1.82	11.33	10.63	10.80
	$\widehat{var}(\hat{y}_2)$	0.23	3.47	-1.25	10.22	10.74	10.28
	$\widehat{var}(\hat{\Delta})$	39.45	11.76	0.76	42.32	15.68	9.09

Table 2: Results from data generated from a multivariate lognormal distribution

		RB			RRMSE		
Correlation		HT	GREG	PROP	HT	GREG	PROP
$\delta_{YY} = 0.2$,	$\hat{y}_1$	-0.20	0.04	-0.14	3.92	4.53	3.84
$\delta_{YX} = 0.2$,	$\hat{y}_2$	0.05	0.18	0.16	3.67	4.41	3.68
$\delta_{XX} = 0.2$.	$\widehat{var}(\hat{y}_1)$	-4.46	-4.55	-6.88	20.93	17.62	20.90
	$\widehat{var}(\hat{y}_2)$	10.24	1.54	3.00	23.50	17.00	20.64
	$\widehat{var}(\hat{\Delta})$	8.15	-9.58	-0.54	25.57	15.59	15.63
$\delta_{YY} = 0.8$,	$\hat{y}_1$	-0.20	0.05	0.00	3.92	2.44	2.36
$\delta_{YX} = 0.8$,	$\hat{y}_2$	-0.07	0.36	0.05	3.68	2.39	2.28
$\delta_{YX} = 0.8$.	$\widehat{var}(\hat{y}_1)$	-4.46	-0.26	-5.97	20.93	16.73	20.08
	$\widehat{var}(\hat{y}_2)$	8.53	3.39	0.43	22.04	16.94	18.78
	$\widehat{var}(\hat{\Delta})$	79.57	-6.58	-3.07	86.55	14.37	15.79

4. CONCLUSIONS

We proposed a multivariate regression estimator that exploits the information from both waves simultaneously instead of each wave separately. This estimator besides using the auxiliary variables, also incorporates the sample design variables.

The simulation study shows that the RRMSE of the proposed point estimator is always smaller than the classical Horvitz-Thompson and generalised regression estimator. With respect to the RRMSE of the variance and variance of the change of the proposed estimator is similar to the other estimator. The variance of change of the proposed estimator has a small relative bias.

REFERENCES

[1] Kalton, G. and Citro, C. F. (1995) Panel surveys: adding the fourth dimension. Innovation: The European Journal of Social Science Research, 8, 25 -39.

[2] Lynn, P. (2009) Methodology of longitudinal surveys. John Wiley & Sons.

[3] Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.

[4] Sarndal, C., Swensson, B. and Wretman, J. (1992) Model assisted survey sampling. Springer-Verlag.

[5] Kish, L. (1965) Survey sampling.

[6] Tam, S. (1984) On covariances from overlapping samples. The American Statistician, 38,

[7] Qualite, L. and Tille, Y. (2008) Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. Survey Methodology, 34, 173 - 181.

[8] Wood, J. (2008) On the covariance between related Horvitz Thompson estimators. Journal of Office Statistics, 24, 53.

[9] Berger, Y. G. and Priam, R. (2015) A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys. To appear in the Journal of Royal Statistical Society, Series A.