Increasing Survey Statistics Precision Using Split Questionnaire Design: An Application of Small Area Estimation

Saeideh Kamgar[1], Hamidreza Navvabpour[1]
[1]Allameh Tabatabaei University, Tehran, Iran
e-mail: saeideh.kamgar@gmail.com

Abstract

A connection between length of a survey questionnaire and the response rate, response burden and precision of survey statistics is an interested topic in survey research methods. Several studies reveal that lengthy survey questionnaires decline the response rates. Split Questionnaire method which introduced as a solution to decrease the non-response rate and response burden, involves splitting the questionnaire into sub-questionnaires and then administering these sub-questionnaires to different subsets of an original sample. As an alternative to this approach we suggest a method of designing and analyzing split questionnaire, using small area estimation. This method relies on the fact that, in the split questionnaire method each sample unit obviously does not respond to all items and consequently, for each item there is not enough sample to support direct estimates of sufficient precision. In a simulation study we show our approach provides more reliable statistics than existed methods.

Keywords: Response burden, matrix sampling, empirical best linear unbiased prediction, multiple imputation.

1. Introduction

Many studies in the Survey Research Methodology are focused on effects of a lengthy survey questionnaire on declining response rate and precision of survey statistics. Split questionnaire has been introduced as a solution to decrease the response burden arising from lengthy questionnaire. This method involves splitting the questionnaire into sub-questionnaires and each one is assigned to a group of sample units. Under a split questionnaire design, procedure of sub-sample selection is at random, therefore, the resulting nonresponse is completely at random.

The literature contains a number of efforts to introduce a method to decrease the length of a given interview using split questionnaire method. This technique has been presented by Raghunathan and Grizzle (1995) as a generalization of multiple matrix sampling design (Shoemaker 1973). They proposed imputation method to fill out the nonresponses from split questionnaire by applying Gibbs sampling under a general location scale model. The origin, applications and outline of current research in this subject has been reviewed in depth by Gonzalez and Eltinge (2007).

Adiguzel and Wedel (2008) proposed a strategy to design the optimal split questionnaire for massive surveys by applying the kullback–Leibler distance. Moreover, the Markov Chain Monte Carlo procedures have been used in order to impute missing values in this method. An optimal split questionnaire design with respect to sample was introduced by Chipperfield and Steel (2009). They also considered estimators (in the simple case of two variables) including the best linear unbiased estimator to impute missing data. Merkouris (2010) suggested an estimation method to improve the precision of survey estimates in matrix sampling survey. The proposed approach was based on the correlation among items of questionnaire. They have used a proper calibration scheme based on the best linear unbiased estimation.

According to the splitting long questionnaire strategies, each sub-questionnaire is asked only from a part of the sample units. Consequently, enough sample units would not be available for each sub-questionnaire.

In this paper, we suggest a method to design and analyze the split questionnaire, using small area estimation technique. We propose small area estimation method as a solution of insufficient sample sizes in split questionnaire method, in order to improve the efficiency of survey statistics. In section 2, we will introduce a new design for split questionnaire which is required to apply small area estimation approach. Section 3 is devoted to describe the use of small area method to estimate population parameters in split questionnaire design. In section 4 we will implement the proposed method and multiple imputation approach on a simulated split questionnaire data. The estimates of population mean, absolute relative bias and mean square error as results of this simulation study are presented in section 5.

2. Split Questionnaire Design

A novel split questionnaire design is required to apply small area estimation. In order to split questionnaire, a new algorithm is proposed as described below:

i. The original questionnaire is divided to (m) sub-questionnaires. Some common items as covariates are assigned to the all sub-questionnaires. Therefore, all sample units respond to them.
ii. All sample units are classified with respect to a known auxiliary variable. Consequently, we make homogeneity within classes in this manner. Each class is considered as an area.
iii. Sample units which belong to each area randomly divided into (m) sub-samples. In each class, each sub-questionnaire is administrated to a sub-sample. Note that in each class, the number of sub-questionnaires and number of sub-samples should be equal.
iv. Step iii is repeated for all classes.

Let $n_i, i = 1, \dots, I$ denotes the sample size in the i^{th} level of a known categorical auxiliary variable where $n = n_1 + n_2 + \cdots + n_I$ is the target sample size. According to the algorithm, $n_i' = \frac{n_i}{m}$, i=1,…,I is the sample size within the i^{th} class devoted to each sub-questionnaire. Hence, instead of n, there are $n' = \frac{n}{m}$ sample units that respond to each sub-questionnaire, where $n' = \sum_{i=1}^{I} n_i'$. The pattern of administering sub-questionnaires to sub-samples is shown in Table 1.

Table 1: Pattern of administering subquestionnaires to sub-samples in small area estimation approach

	Sub-questionnaire / Sub-Samples	*1*	*2*	…	*m*
1	1	✓			
	2		✓		
	⋮	⋮	⋮	⋮	
	m				✓
2	1	✓			
	2		✓		
	⋮	⋮	⋮	⋮	
	m				✓
	⋮	⋮	⋮	⋮	
I	1	✓			
	2		✓		
	⋮	⋮	⋮	⋮	
	m				✓

3. Population characteristics Estimation

As discussed earlier, there is not large enough sample to support direct estimates of appropriate precision based on the proposed design. Accordingly, small area estimates would be useful in this case.

In the case of existing auxiliary information for each unit, one of the common models which has been used in small area estimation is nested error regression model (Rao 2003). This model is a special case of unit level linear mixed model with a block diagonal covariance structure. Under the assumption that population size in the i^{th} area , N_i, is large, the model can be described as:

$$y_{ij} = x_{ij}{}'\boldsymbol{\beta} + v_i + e_{ij},\ j = 1, \dots, n_i{}';\ i = 1, \dots, I;\ n' = \sum_{i=1}^{I} n_i{}' \tag{3.1}$$

Where x_{ij} is a vector of auxiliary variable, y_{ij} the response variable, $n_i{}'$ the sample size in the i^{th} area, $\boldsymbol{\beta}$ is the vector of regression coefficients and v_i is an area-specific random effect with distribution $N(0, {\sigma_v}^2)$. The distribution of error term e_{ij}, is considered as $N(0, {\sigma_e}^2)$ and e_{ij} is assumed to be independent of v_i (Rao 2003).

The empirical best linear unbiased predictor (EBLUP) in the context of linear mixed model is a model-based prediction that can improve the efficiency of small area estimation. The EBLUP is given by

$$\bar{Y}_{EBLUP,i} = \bar{\boldsymbol{X}}_{\boldsymbol{i}}{}'\tilde{\boldsymbol{\beta}} + \hat{\gamma}_i\left(\bar{y}_i - \bar{\boldsymbol{x}}_{\boldsymbol{i}}{}'\tilde{\boldsymbol{\beta}}\right) \tag{3.2}$$

where $\bar{\boldsymbol{X}}_{\boldsymbol{i}}$, $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ and $\bar{y}_i$ are the auxiliary population mean vector, auxiliary sample mean vector and sample-base mean of the i^{th} area, respectively. Furthermore $\tilde{\beta}$ and $\hat{\gamma}_i$ can also be expressed as:

$$\tilde{\beta} = \left(\sum_{i=1}^{K}\sum_{j=1}^{n_i{}'}(x_{ij}\,x_{ij}{}' - \gamma_i\bar{x}_i\bar{x}_i{}')\right)^{-1}\sum_{i=1}^{K}\sum_{j=1}^{n_i{}'}(x_{ij}y_{ij} - \gamma_i\bar{x}_i\bar{y}_i), \tag{3.3}$$

$$\hat{\gamma}_i = \frac{\hat{\sigma}_v{}^2}{\hat{\sigma}_v{}^2 + \hat{\sigma}_e{}^2/n_i{}'}. \tag{3.4}$$

The MSE of estimator (3.2) denoted by MSE $(\bar{Y}_{EBLUP,i}) \approx C_{1,i}{+}C_{2,i} + 2C_{3,i}$, where its components defined as follows:

$$C_{1,i} = \hat{\gamma}_i\left(\hat{\sigma}_e{}^2/n_i{}'\right), \tag{3.5a}$$

$$C_{2,i} = (\bar{X}_i - \hat{\gamma}_i\bar{x}_i)'\left(\hat{\sigma}_e{}^{-2}\textstyle\sum_i\sum_j(x_{ij}x_{ij}{}' - \hat{\gamma}_i n_i{}'\bar{x}_i\bar{x}_i{}')\right)^{-1}(\bar{X}_i - \hat{\gamma}_i\bar{x}_i), \tag{3.5b}$$

$$C_{3,i} = n_i{}'^{-2}\left(\hat{\sigma}_v{}^2 + \hat{\sigma}_e{}^2/n_i{}'\right)^{-3}\hat{\sigma}_e{}^4V_v + \hat{\sigma}_v{}^4V_e - 2\hat{\sigma}_v{}^2\hat{\sigma}_e{}^2V_{ve}. \tag{3.5c}$$

In the above formula V_v and V_e are the asymptotic variances of the estimators $\hat{\sigma}_v{}^2$ and $\hat{\sigma}_e{}^2$, and V_{ve} is the asymptotic covariance of $\hat{\sigma}_v{}^2$ and $\hat{\sigma}_e{}^2$(Rao 2003).

The model (3.1) is used for each sub-questionnaire to compute the EBLUP of population totals for each area. Due to obvious independency of each area form the others, we can use stratified sampling formula for population mean $\bar{Y}_U$. Therefore, the estimate of $\bar{Y}_U$ for the population of size N takes the form

$$\widehat{\bar{Y}_U} = \sum_{i=1}^{I}\left(\frac{N_i}{N}\right)\left(\bar{\boldsymbol{X}}_{\boldsymbol{i}}{}'\tilde{\boldsymbol{\beta}} + \hat{\gamma}_i(\bar{y}_i - \bar{\boldsymbol{x}}_{\boldsymbol{i}}{}'\tilde{\boldsymbol{\beta}})\right) \tag{3.6}$$

and the MSE of $\hat{Y}_U$ can be described as:

$$MSE\left(\widehat{\bar{Y}_U}\right) = \sum_{i=1}^{I}\left(\frac{N_i}{N}\right)^2\left(C_{1,i} + C_{2,i} + 2C_{3,i}\right) \tag{3.7}$$

4. A Simulation Study

This section describes a comparative study of our approach to estimate parameters in surveys with long questionnaire and multiple imputation approach.

For this purpose, we created a questionnaire with 17 questions. Then we spited the questionnaire into five different components based on split questionnaire design (Raghunathan and Grizzle 1995). The first component consists of five items which are highly correlated with other twelve items. These questions were administered to all sample units. This part of questionnaire is a core part of the design. In the rest of the components, three items are assigned to each one in such a way that the within component correlation is small whereas, items in different components are highly correlated .Each double combination of these four components plus the core part compose 6 subquestionnaire, individually.

4.1 Data generator

A multivariate normal random vector is generated 50,000 times, under the correlation pattern described earlier. We also produce a multinomial variable as a stratification variable which is strongly correlated with the other variables. In order to simulate data for the proposed split questionnaire, the population units were classified based on the stratification variable into the five stratums. A simple random sample without replacement of a fixed size n=2000 is selected from the population. Sample units in each stratum were randomly assigned to the all six subquestionnaires. The population mean of each item is estimated by applying multiple imputation approach using the predictive mean matching method (Rubin 1987) and the small area estimation technique. To compare two approaches we generate 1000 simulated bootstrap samples.

4.2 Measures of Comparisons

It is useful to establish some notations before presenting results of the simulation study,

I. Bias. Bias for a parameter estimate is computed by subtracting the true parameter value (θ) from the average of the estimates parameter value ($\hat{\bar{\theta}}$) for the 1000 simulated sample, e.g.

$$Bias(\hat{\theta}) = \hat{\bar{\theta}} - \theta \tag{3.8}$$

where $\hat{\bar{\theta}} = \frac{1}{1000}\sum_{i=1}^{1000} \hat{\theta}_i$ and $\hat{\theta}_i$ is the i^{th} bootstrap estimate.

II. Estimated absolute relative bias (EARB). Absolute relative bias is estimated as:

$$EARB(\hat{\theta}) = \frac{\left|\hat{\bar{\theta}} - \theta\right|}{\theta} \tag{3.9}$$

III. Estimated mean square error (EMSE). The EMSE for estimated parameter is the sum of the estimated true variance and the squared estimated bias, e.g.

$$EMSE(\hat{\theta}) = \widehat{var}(\hat{\theta}) + (\widehat{bias}(\hat{\theta}))^2 \tag{3.10}$$

where $\widehat{var}(\hat{\theta}) = \frac{1}{1000-1}\sum_{i=1}^{1000}\left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2$

IV. Estimated relative efficiency (ERE). The ERE of an estimator based on method 2 relative to method 1 is defined as

$$ERE = \frac{EMSE(\hat{\theta})_{method1}}{EMSE(\hat{\theta})_{method2}} \tag{3.11}$$

5. Results of the Study

Bootstrap population mean estimate (BPME) of each question and corresponding bootstrap estimates of absolute relative bias and mean square error for two approaches (small area using sample auxiliary information and multiple imputation technique) are presented in Table 2. It reveals that small area estimates mostly have lower EARB relative to multiple imputation based estimates. Moreover, there were no cases in which the multiple imputation approach gives a smaller MSE than the small area method across all items. Hence, it seems that small area estimates are more precise than multiple imputation estimates.

We have also applied small area approach to calculate bootstrap estimates of the population means when we use auxiliary information from the population and not from the sample. Similar measures of comparisons are calculated and presented in Table 3.

As expected, the efficiency of small area estimates (using population auxiliary) is still higher than multiple imputation approach. Furthermore, our assessments indicate that Small Area technique requires less computation comparing with multiple imputation method. Moreover, Small Area method does not require to produce data points, hence it would be more applicable, where the goals is to improve survey statistics quality and not to improve survey data quality.

Table 2: Absolute relative bias, MSE and relative efficiency for 1000 bootstrap samples using sample auxiliary information

Item Num.	*Small Area **with** sample auxiliary Info. (Method 1)*			*Multiple Imputation (Method 2)*			*ERE Of Method 2 with respect to Method 1*
	BPME	EARB (%)	EMSE (%)	BPME	EARB (%)	EMSE (%)	
1	11.085	0. 468	0.196	11.084	0.482	0.844	0.232
2	11.080	0. 127	0.097	11.086	0.182	0.704	0.138
3	10.676	0.147	0.118	10.675	0.156	0.642	0.184
4	12.174	0.277	0.120	12.175	0.276	0.772	0.156
5	11.467	0.062	0.155	11.473	0.111	0.777	0.200
6	10.717	0.307	0.164	10.712	0.354	0.731	0.224
7	08.885	0.102	0.104	08.893	0.190	0.639	0.163
8	12.428	0.030	0.242	12.425	0.053	0.954	0.253
9	10.515	0.585	0.160	10.521	0.526	0.737	0.217
10	12.128	0.271	0.300	12.132	0.245	0.857	0.350
11	10.229	0.629	0.286	10.226	0.661	0.860	0.333
12	13.089	0.319	0.397	13.077	0.233	0.973	0.408

Table 3: Absolute relative bias, MSE and relative efficiency for 1000 bootstrap samples using population auxiliary information

Item Num.	*Small Area **with** pop. auxiliary Info (Method 3)*			*Multiple Imputation (Method 2)*			*ERE Of Method 2 with respect to Method 3*
	BPME	EARB (%)	EMSE (%)	BPME	EARB (%)	EMSE (%)	
1	11.105	0.295	0.188	11.084	0.482	0.844	0.222
2	11.098	0.287	0.115	11.086	0.182	0.704	0.163
3	10.690	0.012	0.082	10.675	0.156	0.642	0.128
4	12.194	0.118	0.094	12.175	0.276	0.772	0.122
5	11.490	0.267	0.175	11.473	0.111	0.777	0.226
6	10.730	0.190	0.133	10.712	0.354	0.731	0.182
7	08.901	0.284	0.140	08.893	0.190	0.639	0.219
8	12.450	0.146	0.224	12.425	0.053	0.954	0.234
9	10.530	0.445	0.124	10.521	0.526	0.737	0.169
10	12.146	0.126	0.315	12.132	0.245	0.857	0.368
11	10.245	0.478	0.249	10.226	0.661	0.860	0.290
12	13.111	0.487	0.376	13.077	0.233	0.973	0.386

6. References

Adiguzel, F., Wedel, M. (2008), "Split Questionnaire Design for Massive Surveys," *American Marketing Association,* 45(5): 608-617.

Chipperfield, J. O., and Steel, D. G. (2009), "Design and Estimation for Split Questionnaire Surveys," *Journal of Official statistics,* 25: 227-244.

Gonzalez, J. M., and Eltinge, J. L. (2007), "Multiple Matrix Sampling: A review," *American Statistical Association, Proceedings of the Section on Survey Research Methods,* 3069–3075.

Merkouris, T. (2010), "An Estimation Method for Matrix Survey Sampling," *Section on Survey Research Method.*

Raghunathan, T.E. and Grizzle, J.E. (1995), "A Split Questionnaire Survey Design," *Journal of the American Statistical Association,* 90, 54-63.

Rao, J. N. K. (2003), *Small Area Estimation*, New York: John Wiley& Sons.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley& Sons.

Shoemaker, D. M. (1973), *Principles and Procedures of Multiple Matrix Sampling*, Cambridge, MA: Ballinger.