ATLANTIS PRESS Journal of Statistical Theory and Applications, Vol. 16, No. 4 (December 2017) 467–475

Interactive Linear Models in the Context of Two-Stage Sampling

Pulakesh Maiti
Economic Research Unit, Indian Statistical Institute
203 Barrackpore Trunk Road, Kolkata 700108, India
pulakesh@isical.ac.in

Received 4 September 2014

Accepted 4 March 2015

We study an interactive linear model that incorporates investigator and/or supervisor interventions in the context of a two-stage sampling design. We obtain an unbiased estimator of the finite population total and derive an unbiased estimator of its variance. Our method builds on Sinha and Maiti (2014), which addressed a single-stage cluster sampling design.

Keywords: Finite population inference; Horvitz-Thompson estimator; Two-stage sampling design; SRSWOR; Investigator intervention; Supervisor intervention; Linear model; Variance components, Blinded submission; Unblinded submission

2000 Mathematics Subject Classification: 62D05, 94A20

1. Introduction to Sampling Design, Survey Design and Interactive Linear Model

The general models for variability in the literature are expressed either as a mean squared error decomposition model or as a linear model [Lessler and Kalsbeek (1992)]. The net bias is assumed to be zero, so that the models deal only with variability. The major difference between the two types of models is that the decomposition model approach often has a component attributable to the interaction between sampling and measurement errors, whereas the linear model approach omits this component. Here, we adopt a linear model approach, and we assess a direct response on a quantitative variable Y measured on selected units from a finite, labeled population of size N.

Of course, in actual surveys, we need investigators and/or supervisors as well. We consider a situation wherein there are possibilities of investigators' and/or supervisors' intervention(s) on the response profiles ultimately received by the data collection agency. As in Sinha and Maiti (2014), we assume that these intervention effects are random with mean 0, and they are non-interactive both within and between the data collection and the data handling personnels. The problem is to unbiasedly estimate the finite population total of the response variable Y based on a two-stage sampling design which is administered in a situation wherein the above two types of random interventions are likely to be present. When the target population is not of the type discussed here, the mean squared error decomposition model can be utilized profitably.

1.1. *Sampling design*

To fix ideas, let us assume that all N population units are grouped into M clusters of size L each (that is, $N = ML$). An $SRSWOR(M,m)$ design is applied to the M clusters in the population to select m clusters, followed by $SRSWOR(L,l)$ designs applied simultaneously to all selected clusters. For example, suppose that we have a finite population of $N = 2800$ respondents who are grouped into $M = 70$ clusters of $L = 40$ respondents each. A random sample of $m = 7$ clusters is chosen following a fixed size sampling design, say, for example, an $SRSWOR(70,7)$ design. This is followed by simultaneous (but independent) applications of $SRSWOR(40,10)$ designs to the 7 selected clusters. Thus, altogether 70 ultimate respondents are selected from the 7 clusters (each contributing $l = 10$ units) by a two-stage sampling design.

1.2. *Survey design*

Let there be t investigators and s supervisors engaged in the process of data collection and supervision whose interventions are likely to be present and must be accounted for. We introduce an allocation matrix of order $t \times m$ to describe the assignment of t investigators to the m sampled clusters. We also introduce another allocation matrix of order $s \times t$ to describe the nature of supervisor-investigator 'dual checks' on the cluster-based respondent profiles. While proposing these two types of allocation matrices, we may take recourse to some 'nice' combinatorial structure such as a BIBD. The important point to be noted is that each cluster-based respondent's profile should be collected by more than one investigator and then the profile must be checked by at least one supervisor.

In the illustrative example above, let us label the $m = 7$ selected clusters as I to VII. Further suppose that there are $t = 7$ investigators and $s = 2$ supervisors. The 7 investigators are assigned to the 7 selected clusters using a symmetric $BIBD(7,7,3,3,1)$ generated from the initial set $(1,2,4)$ according to Bose's technique [Bose (1939)]. Supervisor 1 oversees Investigators 1–4, and Supervisor 2 oversees Investigators 4–7. In particular, both supervisors oversee Investigator 4. These allocation matrices are given in the Appendix. The resulting survey design is as follows:

Assigning investigators and supervisors to selected CRUs

CRU i	(j,k) combination of Investigator-Supervisor
I	(1, 1); (5, 2); (7, 2)
II	(1, 1); (2, 1); (6, 2)
III	(2, 1); (3, 1); (7, 2)
IV	(1, 1); (3, 1); (4, 1); (4, 2)
V	(2, 1); (4, 1); (4, 2); (5, 2)
VI	(3, 1); (5, 2); (6, 2)
VII	(4, 1); (4, 2); (6, 2); (7, 2)

Since, the cluster sizes are all equal $(l = 10)$, we ignore the cluster size effect, and treat each cluster as a composite respondent unit (CRU).

1.3. *Data accrued from the field*

Denote by i a CRU in the first-stage sample of size m clusters and by $S[i]$, the number of schedule based observations collected on the particular CRU. Naturally $S[i]$ is based on the survey design used for this CRU in combination with the scheme of involvement of investigators-supervisors. Let

us define $I[i:(j,k)]=1$, if Investigator j and Supervisor k have both worked on collecting and checking information from CRU i; and otherwise, $I[i:(j,k)]=0$. Then, clearly

$$S[i]=\Sigma_j\,\Sigma_k\,I[i:(j,k)]$$

Whenever $I[i:(j,k)]=1$, we denote by $Y_{[i:(j,k)]}$ the underlying (true, but unobserved) total response on the study variable Y from CRU i. Without any intervention effect on the part of the investigators/supervisors and without any measurement errors, the different readings on $Y_{[i:(j,k)]}$ for each cluster i would only vary due to second-stage sampling. But in reality, we are usually faced with possible intervention effects, measurement errors in addition to second-stage sampling variation. So, we denote by $\hat{Y}_{[i:(j,k)]}$ an estimate of $Y_{[i:(j,k)]}$ based on a second-stage sampling scheme obtained by inverstigator-supervisor combination (j,k). Indeed, given that 10 units are selected out of 40 units within CRU i based on a SRSWOR design, the cluster total is estimated by

$$\hat{Y}_{[i:(j,k)]}=4\times\text{the sample total of 10 }Y\text{-values, given that}\{I[i:(j,k)]=1\}$$

This is the important point of departure of the present paper from Sinha and Maiti (2014). Whereas in the case of a single-stage cluster sampling $Y_{[i:(j,k)]}$ are readily available, in the case of a two-stage sampling design we must estimate these quantities by utilizing second-stage sampling designs.

In our illustrative example above, we have altogether 24 data points $\{\hat{Y}_{[i:(j,k)]}:I[i:(j,k)]=1\}$ with cluster-wise frequencies $(3,3,3,4,4,3,4)$. These 24 observations are used to fit the interactive linear model developed in the next subsection.

1.4. *Modelling $\hat{Y}_{[i:(j,k)]}$ whenever $\{I[i:(j,k)]=1\}$*

Our primary objective is to examine the effects of interventions by one or the other or both sources of intervention, and also account for measurement error in order to provide a valid estimate of the population total along with its standard error, based on two-stage sampling. Towards that end, let us postulate a linear model for $\hat{Y}_{[I:(1,1)]}$. Similar models also apply for all other $\hat{Y}_{[i:(j,k)]}$.

$$\hat{Y}_{[I:(1,1)]}=\hat{T}R_I+IR_1+IS_1+e_{[I:(1,1)]} \tag{1.1}$$

where $\hat{T}R_I$ is the estimate of true total response TR_I of CRU I based on two-stage sampling, IR_1 is the intervention effect of the Investigator 1, IS_1 is that of the Supervisor 1, and the last term is the error. We assume that the error and the intervention effects are all randomly distributed with means 0 and variances $\sigma_e^2,\sigma_{IR}^2,\sigma_{IS}^2$ respectively, while all pairwise effects/interventions are uncorrelated.

Let $\hat{Y}_{i..}$ denote the (simple) average of the $S[i]$ estimates of total Y-value corresponding to CRU i submitted to the agency by various investigator-supervisor combinations. That is,

$$\hat{Y}_{i..}=\frac{\Sigma_j\,\Sigma_k\hat{Y}_{[i:(j,k)]}\,I[i:(j,k)]}{S[i]} \tag{1.2}$$

In view of the model assumptions, we have

$$E_M\left(\hat{Y}_{i..}\right)=\hat{T}R_i;\ \text{ for } i=1,2,\ldots,7 \tag{1.3}$$

where E_M is the expectation under the adopted model. Furthermore, due to second stage sampling,

$$E_2\left(\hat{T}R_i\right)=TR_i;\ \text{ for } i=1,2,\ldots,7 \tag{1.4}$$

where E_2 is the expectation under second stage sampling.

2. Estimator of Population Total T, its Expectation and Variance

We use the conventional Horvitz-Thompson estimator (HTE) to obtain an unbiased estimate of the population total $T(TR)$, the total true response under the two-stage sampling scheme; that is,

$$\hat{T}(TR) = \sum_{i \in s} \frac{\hat{Y}_{i..}}{\pi_i}, \tag{2.1}$$

where π_i is the inclusion probability of CRU i during first-stage sampling. Let E_S, E_1, E_2 and $E_{2|1}$ denote respectively the expectations with respect to overall sampling design, the first- and the second-stage sampling designs, and the conditional expectation with respect to the second-stage sampling design given the first-stage sampling design. Then using (1.3) and (1.4), we have

$$\begin{aligned} E(\hat{T}(TR)) &= E_S E_M \left(\sum_{i \in s} \frac{\hat{Y}_{i..}}{\pi_i} \right) \\ &= E_1 E_{2|1} \left(\sum_{i \in s} \frac{\hat{T}R_i}{\pi_i} \right) \\ &= E_1 \left(\sum_{i \in s} \frac{TR_i}{\pi_i} \right) = T(TR) \end{aligned}$$

Furthermore, the total within CRU i is estimated based on the second-stage sample as

$$\hat{T}R_i = \sum_{h|i} \frac{TR_{hi}}{\pi_{h|i}} \tag{2.2}$$

where TR_{hi} is the true Y-value of the h-th unit in the second-stage sample from CRU i, and $\pi_{h|i}$ is the inclusion probability of that unit in the second-stage sample.

Remark 2.1 In the above (2.1) and (2.2) represent the most general expressions of the estimators of population total and the CRU totals for arbitrary first-stage and second-stage sampling designs.

Let us now study the variance of the estimator of population total given in (2.1). Let V_S, V_1, V_2 and V_M denote respectively the variances with respect to overall sampling design, the first- and the second-stage sampling designs, and the interactive linear model. Also let Cov_M denote the model covariance. Then, under any fixed sample size design [see, for example, Hedayat and Sinha (1991)], we have

$$V(\hat{T}(TR)) = V_S E_M \left(\sum_{i \in s} \frac{\hat{Y}_{i..}}{\pi_i} \right) + E_S V_M \left(\sum_{i \in s} \frac{\hat{Y}_{i..}}{\pi_i} \right) \tag{2.3}$$

Simplifying the first term on the right hand side of (2.3), and using (1.3) and (1.4), we have

$$\begin{aligned} V_S E_M\left(\sum_{i\in s}\frac{\hat{Y}_{i..}}{\pi_i}\right) &= V_S\left(\sum_{i\in s}\frac{\hat{T}R_i}{\pi_i}\right) \\ &= V_1E_2\left(\sum\frac{\hat{T}R_i}{\pi_i}\right)+E_1V_2\left(\sum\frac{\hat{T}R_i}{\pi_i}\right) \\ &= V_1\left(\sum\frac{TR_i}{\pi_i}\right)+E_1\left(\sum\frac{V_2(\hat{T}R_i)}{\pi_i^2}\right) \\ &= \sum_{i<i'}\sum\left(\frac{TR_i}{\pi_i}-\frac{TR_{i'}}{\pi_{i'}}\right)^2(\pi_i\pi_{i'}-\pi_{ii'})+\sum\frac{V_2(\hat{T}R_i)}{\pi_i} \end{aligned} \tag{2.4}$$

where $\pi_{ii'}$ denotes the joint inclusion probability of CRU i and CRU i' in the first-stage sample. The last sum on the right hand side of (2.4) involves summands with numerators of the form

$$V_2\left(\hat{T}R_i\right)\Big) = \sum_{h<h'}\sum\left(\frac{TR_{hi}}{\pi_{h|i}}-\frac{TR_{h'i}}{\pi_{h'|i}}\right)^2\left(\pi_{h|i}\pi_{h'|i}-\pi_{hh'|i}\right)$$

where $\pi_{hh'|i}$ denotes the joint inclusion probability of units h and h' in the second-stage sample from CRU i. Note that $V_2\left(\hat{T}R_i\right)\Big)$ can be unbiasedly estimated by

$$v_2\left(\hat{T}R_i\right)\Big) = \sum_{h<h'}\sum\left(\frac{TR_{hi}}{\pi_{h|i}}-\frac{TR_{h'i}}{\pi_{h'|i}}\right)^2\left(\frac{\pi_{h|i}\pi_{h'|i}-\pi_{hh'|i}}{\pi_{hh'|i}}\right) \tag{2.5}$$

Likewise, the second term on the right hand side of (2.3) evaluates as

$$\begin{aligned} E_S V_M\left(\sum_{i\in s}\frac{\hat{Y}_{i..}}{\pi_i}\right) &= E_S\left[\sum V_M\left(\frac{\hat{Y}_{i..}}{\pi_i}\right)+\sum_{i\neq i'}\sum\mathrm{Cov}_M\left(\frac{\hat{Y}_{i..}}{\pi_i},\frac{\hat{Y}_{i'..}}{\pi_{i'}}\right)\right] \\ &= E_S\left[\sum\frac{V_M(\hat{Y}_{i..})}{\pi_i^2}+\sum_{i\neq i'}\sum\frac{\mathrm{Cov}_M(\hat{Y}_{i..},\hat{Y}_{i'...})}{\pi_i\,\pi_{i'}}\right] \end{aligned} \tag{2.6}$$

The model variances and covariances in (2.6) are computed by pre- and post-multiplying the covariance matrices (given in Section 3) between vectors (of length $S[i]$ and $S[i']$) of all estimates of totals within CRU i and CRU i' obtained by respective (j,k) combinatons, by suitable coefficient vectors (determined by the adopted convention) given in Section 4.

3. Computations of Model-based Variances and Covariances

Since every CRU is viewed as a cluster of 10 second-stage units chosen from 40 availabule units, the response on each 'cluster unit' is four times the sum of the responses of the constituent members of the second-stage sample. Further, since there are three data points for the responses within CRU I (as collected independently by the three investigators 1, 5, 7), we take the average of the three responses and use this as a representative figure, denoted by $\hat{Y}_{I..}$. This we do for all other CRUs as well. In view of the model assumptions, $E_M(\hat{Y}_{i..}) = \hat{T}R_i$; for $i = 1,2,\ldots,7$. Note that, because these 10 units are selected according to a fixed-size sampling design $SRSWOR(40,10)$, $\hat{T}R_i$ is random.

Whenever investigator and/or supervisor interventions are likely to be present and are to be accounted for, investigators and supervisors are allocated to the sampled cluster units as described

in Subsection 1.2. Since the two allocation matrices (for investigators and supervisors) are chosen in advance, the variance and the covariance computations underlying the interaction model will remain the same for all choices of the m clusters, except for their identification in terms of cluster labels. For example, if the randomly selected respondent clusters are labeled [3, 17, 33, 41, 57, 63, 69], then the above expressions for model-based variances and covariances correspond to the cluster in the order mentioned. In other words, $\sum_{I,I}$ in effect corresponds to $\sum_{3,3}$; $\sum_{II,II}$ corresponds to $\sum_{17,17}$; $\sum_{I,II}$ corresponds to $\sum_{3,17}$; and so on. The realized cluster labels in ascending order replace the labels $I, II, \ldots, VII$.

Computations of the model-based variances and covariances are quite laborious (though straight-forward). These are already developed in Sinha and Maiti (2014). As an illustration, we show $\sum_{I,II}$, which stands for the 3×3 covariance matrix between the two random vectors $\hat{Y}_{[I]}$ and $\hat{Y}_{[II]}$ each of length 3 (because three (j,k) combinations worked on CRU I and three on CRU II):

$$\sum_{I,II} = \mathrm{Cov}_M\left(\hat{Y}_{[I]}, \hat{Y}_{[II]}\right) = \begin{bmatrix} \sigma_{IR}^2 + \sigma_{IS}^2 & \sigma_{IS}^2 & 0 \\ 0 & 0 & \sigma_{IS}^2 \\ 0 & 0 & \sigma_{IS}^2 \end{bmatrix}$$

Likewise, we obtain all other $\sum_{i,i'} = \mathrm{Cov}_M\left(\hat{Y}_{[i]}, \hat{Y}_{[i']}\right)$.

The differences between the single-stage cluster sampling design and the two-stage sampling design in the context of interactive linear model is described fully in the Appendix.

4. Data Analysis

We will now discuss the essential features of data analysis for unbiased estimation of the finite population total $T(TR)$ of the study variable Y under the above interactive linear model. We need to differentiate between two distinct scenarios as was done in Sinha and Maiti (2014):

(i) Blinded Submission
(ii) Unblinded Submission

The above two cases refer to the different scenarios of submitting the response profiles to the supervisors. In the blinded case, each supervisor treats each response profile as a separate, isolated document without the knowledge of which CRU it was reorded on and by whom. In the unblinded case, the supervisor receives also the identity of the CRU and the interviewer/investigator along with the response profiles. We will study both scenarios in this paper.

4.1. *Data analysis under blinded submission*

In a very general set up, we consider a finite, labeled population of N units; and we take recourse to a two-stage design as mentioned earlier in Section 1. Because of possible investigator and/or supervisor interventions, the response may be distorted; and we account for this distortion by stipulating an interactive linear model described in (1.1). In view of (1.3), a model-unbiased estimation of cluster total is ensured. Thereafter, we use the conventional Horvitz-Thompson estimator of the population total under two-stage sampling design as mentioned in (2.1). An expression for the variance of $\hat{T}(TR)$ has been given in (2.3)–(2.6). In the next paragraph, we discuss how to estimate the variance of $\hat{T}(TR)$.

The exact computation of $E_S V_M(\bullet)$ is quite involved. However, an unbiased, sampled CRU based estimator of $E_S V_M(\bullet)$ is simply given by $V_M(\bullet)$, assuming that the three variance components $\sigma^2_{IS}, \sigma^2_{IR}, \sigma^2_e$ are known. Likewise, computing an unbiased estimate of $V_S E_M(\bullet)$, by substituting (2.5) in (2.4), would be trivial if the true values of the sampled cluster totals were known. But since the true values are not known, one can make use of $\hat{Y}_{i..}$ in place of TR_i.

Finally, the coefficient vector corresponding to CRU i is of the form $(1,1,\ldots,1)/S[i]$, and is of length $S[i]$ (because the simple average $\hat{Y}_{i..}$ gives equal weight to all $S[i]$ components). Thus, under the assumption of known variance components, we construct the variance estimator.

4.2. *Data analysis under unblinded submission*

Suppose, for example, that $I[i;(j,k)] = I[i;(j',k)] = 1$. That is, Supervisor k has to handle two separate response profiles of the same CRU i submitted by Investigators j and j'. Therefore, under unblinded submission, it makes good sense for Supervisor k first to average out these two measurements on the same CRU i, and then provide his/her own 'input' to that average before forwarding the response to the agency! In constrast, under blinded submission, the supervisor's input was incorporated separately for each response profile submitted to him/her, since the CRU identity was not known to the supervisor.

Continuing the same example as before, but now assuming unblinded submission, we apply the aforementioned scheme of 'averaging' multiple responses from the same CRU as follows:

$$
\begin{aligned}
\hat{Y}_{I**} &= \left[\hat{Y}_{I:(1,1)} + \{\hat{Y}_{I:(5,2)} + \hat{Y}_{I:(7,2)}\}/2\right]/2 = \left[2\hat{Y}_{I:(1,1)} + \hat{Y}_{I:(5,2)} + \hat{Y}_{I:(7,2)}\right]/4 \\
\hat{Y}_{II**} &= \left[\{\hat{Y}_{II:(1,1)} + \hat{Y}_{II:(2,1)}\}/2 + \hat{Y}_{II:(6,2)}\right]/2 = \left[\hat{Y}_{II:(1,1)} + \hat{Y}_{II:(2,1)} + 2\hat{Y}_{II:(6,2)}\right]/4 \\
\hat{Y}_{III**} &= \left[\{\hat{Y}_{III:(2,1)} + \hat{Y}_{III:(3,1)}\}/2 + \hat{Y}_{III:(7,2)}\right]/2 = \left[\hat{Y}_{III:(2,1)} + \hat{Y}_{III:(3,1)} + 2\hat{Y}_{III:(7,2)}\right]/4 \\
\hat{Y}_{IV**} &= \left[\{\hat{Y}_{IV:(1,1)} + \hat{Y}_{IV:(3,1)} + \hat{Y}_{IV:(4,1)}\}/3 + \hat{Y}_{IV:(4,2)}\right]/2 \\
&= \left[\hat{Y}_{IV:(1,1)}) + \hat{Y}_{IV:(3,1)} + \hat{Y}_{VI:(4,1)} + 3\hat{Y}_{IV:(4,2)}\right]/6 \\
\hat{Y}_{V**} &= \left[\{\hat{Y}_{V:(2,1)} + \hat{Y}_{V:(4,1)}\}/2 + \{\hat{Y}_{V:(4,2)} + \hat{Y}_{V:(5,2)}\}/2\right]/2 \\
&= \left[\hat{Y}_{V:(2,1)} + \hat{Y}_{V:(4,1)} + \hat{Y}_{V:(4,2)} + \hat{Y}_{V:(5,2)}\right]/4 \\
\hat{Y}_{VI**} &= \left[\hat{Y}_{VI:(3,1)} + \{\hat{Y}_{VI:(5,2)} + \hat{Y}_{VI:(6,2)}\}/2\right]/2 = \left[2\hat{Y}_{VI:(3,1)} + \hat{Y}_{VI:(5,2)} + \hat{Y}_{VI:(6,2)}\right]/4 \\
\hat{Y}_{VII**} &= \left[\hat{Y}_{VII:(4,1)} + \{\hat{Y}_{VII:(4,2)} + \hat{Y}_{VII:(6,2)} + \hat{Y}_{VII:(7,2)}\}/3\right]/2 \\
&= \left[3\hat{Y}_{VII:(4,1)} + \hat{Y}_{VII:(4,2)} + \hat{Y}_{VII:(6,2)} + \hat{Y}_{VII:(7,2)}\right]/6
\end{aligned}
$$

Essentially, the simple averages $\hat{Y}_{i..}$ in the case of blinded submission are replaced by the weighted averages $\hat{Y}_{i**}$ in the case of unblinded submission. Next, we need to compute the variances and the covariances of the resulting input averages from all seven CRUs. This was considered earlier in Sinha and Maiti (2014), and the expressions for the variances and the covariances were derived. Indeed, the coefficient vectors needed to evaluate the variances and covariances corresponding to CRU i (of length $S[i]$) and CRU i' (of length $S[i']$) are readily seen from the above expressions of weighted averages $\hat{Y}_{i**}$.

As we see it, the non-trivial problems are those of variance estimation in such models with/without two-stage sampling design. We hope to venture further in this area of research.

Remark 4.1 Extension to a two-stage design calls for natural generalization of results in a single stage design, which we have provided here.

Acknowledgements

The author acknowledges Prof. Bikas K. Sinha for his valuable suggestions in preparing the present manuscript.

References

[1] R.C. Bose, On the construction of balanced incomplete block designs, *Ann. Eugen.*, **9**, (1939) 353–399.
[2] A.S. Hedayat and Bikas K. Sinha, *Design and Inference in Finite Population Sampling*, (Wiley, New York, 1991).
[3] Judith T. Lessler and William D. Kalsbeek, *Non-Sampling Errors in Surveys*, (Wiley, New York, 1992).
[4] Bikas K. Sinha and Pulakesh Maiti, Interactive linear models in survey sampling, *Journal of Statistical Theory and Applications*, **13(3)**, (2014) 263–272.

Appendix A. Appendix: Details of Sampling, Interventions, Modelling and Analysis

A.1. *Sampling design*

(i) The population consists of $N = ML$ units, grouped into M clustes of L units each. An $SRSWOR(M,m)$ design is used to select m clusters out of M clusters in Stage 1, followed by $SRSWOR(L,l)$ designs to select l units out of L units within each selected clusters in Stage 2, resulting in $n = ml$ ultimate respondents.

(ii) The data consist of m selected clusters, and $S[i]$ estimates of total Y-response from CRU i collected by the investigators, modified by the supervisors and received by the agency.

A.2. *Survey design*

To each selected CRU, investigators and supervisor(s) are assigned using Tables 1 and 2. The investigator-supervisor assignments within each selected CRU is shown in Subsection 1.3.

Table 1. Assigning investigators to selected CRUs based on $BIBD(7,7,3,3,1)$

CRU / Invest.	I	II	III	IV	V	VI	VII
1	*	*		*			
2		*	*		*		
3			*	*		*	
4				*	*		*
5	*				*	*	
6		*				*	*
7	*		*				*

Table 2. Assigning two supervisors to oversee seven investigators

Invest. / Super.	1	2	3	4	5	6	7
1	*	*	*	*			
2				*	*	*	*

A.3. ***Estimated total Y-response based on second-stage sample from CRU i***

$\hat{Y}_{[i:(j,k)]} = L \cdot \bar{y}_{[i:(j,k)]}$, where $\bar{y}_{[i:(j,k)]}$ is the second-stage sample average obtained after interventions by Investigator j and Supervisor k; and L is the number of second-stage units within each CRU i.

A.4. ***Modelling***

Model under Single-Stage Sampling:

$$Y_{[i:(j,k)]} = TR_i + IR_j + IS_k + e_{[i:(j,k)]}$$

Model under Two-Stage Sampling:

$$\hat{Y}_{[i:(j,k)]} = \hat{T}R_i + IR_j + IS_k + e_{[i:(j,k)]}$$

A.5. ***Estimated TR within CRU i***

Under Single-Stage-Sampling:

$$\hat{T}R_i = \frac{\sum_j \sum_k Y_{[i:(j,k)]} I[i:(j,k)]}{\sum_j \sum_k I[i:(j,k)]}$$

Under Two-Stage Sampling:

$$\hat{T}R_i = \frac{\sum_i \sum_k \hat{Y}_{[i:(j,k)]} I[i:(j,k)]}{\sum_j \sum_k I[i:(j,k)]}$$

A.6. ***Expectation of estimated TR within CRU i reported by (j,k) combination***

$E\left(\hat{Y}_{[i:(j,k)]}\right) = E_S E_M\left(\hat{Y}_{[i:(j,k)]}\right) = E_M E_S\left(\hat{Y}_{[i:(j,k)]}\right) = E_M\left(Y_{[i:(j,k)]}\right) = TR_i$

A.7. ***Expectation of estimated TR within CRU i***

$E\left(\hat{T}R_i\right) = E_S E_M\left(\hat{T}R_i\right) = E_M E_S\left(\hat{T}R_i\right) = E_M\left(Y_{i..}\right) = TR_i$

A.8. ***Variance of estimated TR within CRU i reported by (j,k) combination***

$$\begin{aligned} V_M\left(\hat{Y}_{[i:(j,k)]}\right) &= \sigma^2_{IR} + \sigma^2_{IS} + \sigma^2_e \\ V\left(\hat{Y}_{[i:(j,k)]}\right) &= V_S E_M\left(\hat{Y}_{i[j,k]}\right) + E_S V_M\left(\hat{Y}_{[i:(j,k)]}\right) \\ &= V_S\left(\hat{T}R_i\right) + E_S\left(\sigma^2_{IR} + \sigma^2_{IS} + \sigma^2_e\right) \\ &= V_S\left(\hat{T}R_i\right) + \left(\sigma^2_{IR} + \sigma^2_S + \sigma^2_e\right) \end{aligned}$$

A.9. ***Variance and covariance of estimated TR within CRU i***

Under Single-Stage-Sampling:

The expressions for $V\left(\hat{T}R_i\right)$ and $\mathrm{Cov}\left(\hat{T}R_i, \hat{T}R_{i'}\right)$ are worked out in Sinha and Maiti (2014).

Under Two-Stage-Sampling:

$$V\left(\hat{T}R_i\right) = V_S E_M(\cdot) + E_S V_M(\cdot)$$

The first term on the right hand side, $V_S E_M(\cdot)$, reduces to the variance in a two-stage sampling, and the second term, $E_S V_M(\cdot)$, is the same as in Sinha and Maiti (2014).