Journal of Statistical Theory and Applications, Vol. 13, No. 3 (September 2014), 263-272

Interactive Linear Models in Survey Sampling

Bikas K. Sinha[1] and Pulakes Maiti[2]
[1] *Retired Professor of Statistics, Indian Statistical Institute, Kolkata-700108, India*
[2] *Economic Research Unit, Indian Statistical Institute, Kolkata-700108, India*

Received 10 November 2013
Accepted 28 May 2014

Considered is a linear 'interactive' model in the context of survey sampling. This situation arises when investigator and/or supervisor interventions are contemplated in the responses. An appropriate linear model is introduced to represent the response profile(s) arising out of each respondent-cum-investigator-cum-supervisor combination as per the planned 'design layout'. Two situations [blinded and unblinded submission of responses] are differentiated and corresponding data analysis techniques are discussed. Variance components are assumed to be known in this study.

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

1. Introduction to Survey Design and Interactive Linear Model

Considered is the set-up of simple i.e., direct response on a quantitative response variable Y in the context of a finite labeled population of size N. It so happens that in actual surveys, we need investigators and often some supervisors as well. The instruction manuals are prepared for the investigators to maintain uniformity in data collection. The field level data are collected by the investigators. The scrutiny manual is prepared for scrutiny of the filled-in schedules by the supervisors. This is accomplished independently of the investigators. We depict a situation wherein there are possibilities of investigator intervention effect and/or supervisor intervention effect on the response profiles before the same are finally received by the data collection agency. Of course, these intervention effects may be assumed to be random, having mean zero, non-interactive within and between the two sets of 'people'. The problem is to unbiasedly estimate the finite population total of the response variable Y by incorporating a fixed size (n) sampling design and by administering the sampling design in a situation wherein the above two types of random effects are likely to be present.

Denote by i a respondent unit in the sample of size n and by $S[i]$ the number of schedule-based observations collected on this particular unit. It is quite possible that a respondent unit is composed of more than one individual. In this article, we will deal with fixed-size non-overlapping clusters of individuals to represent such respondent units. These clusters are formed before the sampling

operation takes place and the investigators/supervisors are supposed to provide information in the form of sub-totals, accrued from each member of the selected clusters. We will refer to such clusters as Respondent Clusters [RC] and these will be treated as the responding units in the finite population under consideration. The RC subtotals are denoted by the generic notation 'Y'. Naturally, $S[i]$, based on the RC labelled i, is the number of all field-based sub-totals used for this RC in combination with the investigators and the supervisors. We may write $S[i] = \sum\sum I[i;(j,k)]$ where $I[i;(j,k)] = 1$ if (j,k)-combination of the investigator and the supervisor have both worked on a schedule assigned to the ith RC. Naturally, for any triplet $[i;(j,k)]$, $I[i;(j,k)] \geqslant 0$ while $S[i] > 0$ for each responding unit. Whenever $I[i;(j,k)] = 1$, we will denote by $Y_{[i;(j,k)]}$ the underlying response on the study variable for the RC labelled i.

To fix ideas, let us take up the following simple example of a study design involving $n = 7$ RCs selected out of a large number of $N = 70$ RCs, following a fixed size (n) sampling design, say, for example, SRSWOR of 7 RCs, each RC being of size 10. Let there be 7 investigators and 2 supervisors engaged in the process. We designate the RCs as RCI to $RCVII$. Here is the description of a study design [as against the sampling design specified above]:

Choices of $[i;\ (j,k)]$ for which $I[i;\ (j,k)] = 1$

(RCI): $(j = 1;\ k = 1);\ (j = 5,\ k = 2);\ (j = 7;\ k = 2);$
(RCII): $(j = 1,\ k = 1);\ (j = 2,\ k = 1);\ (j = 6,\ k = 2);$
(RCIII): $(j = 2,\ k = 1);\ (j = 3,\ k = 1);\ (j = 7,\ k = 2);$
(RCIV): $(j = 1,\ k = 1);\ (j = 3,\ k = 1);\ (j = 4,\ k = 1);\ (j = 4,\ k = 2);$
(RCV): $(j = 2,\ k = 1);\ (j = 4,\ k = 1);\ (j = 4,\ k = 2);\ (j = 5,\ k = 2);$
(RCVI): $(j = 3,\ k = 1);\ (j = 5,\ k = 2);\ (j = 6,\ k = 2);$
(RCVII): $(j = 4,\ k = 1);\ (j = 4,\ k = 2);\ (j = 6,\ k = 2);\ (j = 7,\ k = 2);$

This study design is essentially derived from a symmetric BIBD(7,7,3,3,1) 'developed from the initial set' $(1,2,4)$, following Bose's technique. Vide Raghavarao (1971). In essence, the above allocation design suggests that the first $RC(I)$ will be approached once by the investigator number 1 and the response profile will be checked by the supervisor 1. Further, the same $RC(I)$ will also be approached by investigators 5 and 7 and both the profiles will be checked by supervisor number 2. Similar explanation applies to the other selected RCs as well. Since the RC sizes are all equal (= 10, in the above), we will ignore the RC size effect and treat each one as a singleton.

Let us denote by $Y_{[I]}, Y_{[II]}, \ldots, Y_{[VII]}$ the 'data' accrued from the field. Without any intervention effect on the part of the investigators/supervisors, we would have regarded the above data as 'error-free' and so usual estimation techniques could be routinely used. Note that in such 'error-free' scenario, there is no difference between $Y_{[I;(1,1)]}$, $Y_{[I;(5,2)]}$ and $Y_{[I;(7,2)]}$, for example. However, we want to examine the possibility of intervention by one or the other group or possibly by both and so we postulate a linear model of the following form, as applied to $Y_{[I;(1,1)]}$, for example:

$Y_{[I;(1,1)]} = TR_{[I]} + IR_1 + S_1 + e_{[I;(1,1)]}$
where $TR_{[I]}$ is the true response from Cluster I, IR_1 is the intervention effect of Investigator 1 and S_1 is that of the Supervisor 1. The last term is the so-called error term. As usual, we assume that the errors and the intervention effects are all randomly distributed with means 0's, variances σ_e^2, σ_{IR}^2, σ_S^2 respectively while all pairwise effects / interventions are uncorrelated. We refer to Searle (1971) for basic results in linear models.

At this stage, we need to differentiate between two distinct scenarios:

(i) Blinded Submission;
(ii) Unblinded Submission.

The above refers to the submission of the response profiles to the supervisors. In case it is blinded, each supervisor treats each response profile as a separate document and treats it as an isolated document - without the knowledge of identification of the interviewer/investigator. In the other case, the supervisor also receives information about the identity of the interviewer/investigator along with response profiles. We will treat both the scenarios in this paper.

2. Interactive Linear Model under Blinded Submission

In the above, since every respondent unit (RC) is viewed as a cluster of 10 units, the response on each RC is taken to be the sum of the responses of the constituent members. Further, since there are 3 data points for the first set i.e., *RCI* – as collected independently by the investigators 1, 5, 7, we straightaway take the average of the three responses and use this as the representative figure for the first selected RC. This we do for all other RCs as well. Note that there are altogether 24 data points in the above study design and the RC-wise frequency distributions are given by 3, 3, 3, 4, 4, 3, 4 respectively. We denote by Y the row vector of 24 observations represented in the order these are displayed above through (RCI) to $(RCVII)$ and by $\mathbf{A}$ the 7×24 incidence matrix of the population units versus the observations as per the sampling design. Thus, for example, the first row vector of $\mathbf{A}$ is given by: $(1\ \ 1\ \ 1\ \ 0\ \ 0\ldots)$. Also we denote by $\mathbf{Y}_{i..}$ the average of the sample observations corresponding to the RC(i) in the sample for $i = I,\ II, \ldots,\ VII$. In view of the model assumptions, $\mathrm{E}_M \mathbf{Y}_{i..} = TR_{[i]};\ i = 1, 2, \ldots, 7$. Computations of the model-based variances and covariances are quite involved and these are developed below.

$$\begin{aligned}\Sigma_{11} =&\ \text{dispersion matrix of } Y_{[I;(1,1)]};\ Y_{I;(5,2)]};\ Y_{[I;(7,2)]}\\ =&\ \text{dispersion matrix of } (IR_1 + S_1 + e_{[I:(1,1)]};\ IR_5 + S_2 + e_{[I;(5,2)]};\ IR_7 + S_2 + e_{[I;(7,2)]})\\ =&\ [(\sigma_e^2 + \sigma_{IR}^2 + \sigma_S^2, 0, 0); (0, \sigma_e^2 + \sigma_{IR}^2 + \sigma_S^2, \sigma_S^2); (0, \sigma_S^2, \sigma_e^2 + \sigma_{IR}^2 + \sigma_S^2)].\end{aligned}$$

Therefore, $\mathrm{V}_M \mathbf{Y}_{I..} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 5\sigma_S^2]/9$.
Likewise,
$\mathrm{V}_M \mathbf{Y}_{II..} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 5\sigma_S^2]/9$;
$\mathrm{V}_M \mathbf{Y}_{III..} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 5\sigma_S^2]/9$;
$\mathrm{V}_M \mathbf{Y}_{IV..} = [4\sigma_e^2 + 6\sigma_{IR}^2 + 10\sigma_S^2]/16$;
$\mathrm{V}_M \mathbf{Y}_{V..} = [4\sigma_e^2 + 6\sigma_{IR}^2 + 8\sigma_S^2]/16$;
$\mathrm{V}_M \mathbf{Y}_{VI..} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 5\sigma_S^2]/9$;
$\mathrm{V}_M \mathbf{Y}_{VII..} = [4\sigma_e^2 + 6\sigma_{IR}^2 + 10\sigma_S^2]/16$.

Next note that the rows of the matrix $\mathbf{A}$ have been numbered as 1 to 24 in a way that these have 1:1 correspondence with the triplets designated for *RCI* to *RCVII*. This representation is already mentioned above.

We now work out Σ_{12} which is a 3×3 vector and stands for the covariance between the two vectors $\mathrm{Y}_{[I]}$ and $\mathrm{Y}_{[II]}$. This is given by $\Sigma_{12} = \mathrm{Cov}_M(\mathrm{Y}_{[I]}, \mathrm{Y}_{[II]}) = [(\sigma_{IR}^2 + \sigma_s^2, \sigma_s^2, 0); (0, 0, \sigma_s^2); (0, 0, \sigma_s^2)]$.

Similarly, we may deduce the following expressions for other covariance matrices:

$\Sigma_{13} = [(\sigma_s^2, \sigma_s^2, 0); (0, 0, \sigma_s^2); (0, 0, \sigma_{IR}^2 + \sigma_s^2)]$,
$\Sigma_{14} = [(\sigma_{IR}^2 + \sigma_s^2, \sigma_s^2, \sigma_s^2, 0); (0, 0, 0, \sigma_s^2); (0, 0, 0, \sigma_s^2)]$,

$\Sigma_{15} = [(\sigma_s^2,\sigma_s^2,0,0);(0,0,\sigma_s^2,\sigma_{IR}^2+\sigma_s^2);(0,0,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{16} = [(\sigma_s^2,0,0);(0,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2);(0,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{17} = [(\sigma_s^2,0,0,0);(0,\sigma_s^2,\sigma_s^2,\sigma_s^2);(0,\sigma_s^2,\sigma_s^2,\sigma_{IR}^2+\sigma_s^2)]$,
$\Sigma_{23} = [(\sigma_s^2,\sigma_s^2,0);(\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,0);(0,0,\sigma_s^2)]$,
$\Sigma_{24} = [(\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,\sigma_s^2,0);(\sigma_s^2,\sigma_s^2,\sigma_s^2,0);(0,0,0,\sigma_s^2)]$,
$\Sigma_{25} = [(\sigma_s^2,\sigma_s^2,0,0);(\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,0,0);(0,0,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{26} = [(\sigma_s^2,0,0);(\sigma_s^2,0,0);(0,\sigma_s^2,\sigma_{IR}^2+\sigma_s^2)]$,
$\Sigma_{27} = [(\sigma_s^2,0,0,0);(\sigma_s^2,0,0,0);(0,\sigma_s^2,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{34} = [(\sigma_s^2,\sigma_s^2,\sigma_s^2,0);(\sigma_s^2,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,0);(0,0,0,\sigma_s^2)]$,
$\Sigma_{35} = [(\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,0,0);(\sigma_s^2,\sigma_s^2,0,0);(0,0,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{36} = [(\sigma_s^2,0,0);(\sigma_{IR}^2+\sigma_s^2,0,0);(0,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{37} = [(\sigma_s^2,0,0,0);(\sigma_s^2,0,0,0);(0,\sigma_s^2,\sigma_s^2,\sigma_{IR}^2+\sigma_s^2)]$,
$\Sigma_{45} = [(\sigma_s^2,\sigma_s^2,0,0);(\sigma_s^2,\sigma_s^2,0,0);(\sigma_s^2,\sigma_{IR}^2+\sigma_s^2,\sigma_{IR}^2,0);(0,\sigma_{IR}^2,\sigma_{IR}^2+\sigma_s^2),\sigma_s^2]$,
$\Sigma_{46} = [(\sigma_s^2,0,0);(\sigma_{IR}^2+\sigma_s^2,0,0);(\sigma_s^2,0,0);(0,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{47} = [(\sigma_s^2,0,0,0);(\sigma_s^2,0,0,0);(\sigma_{IR}^2+\sigma_s^2,\sigma_{IR}^2,0,0);(\sigma_{IR}^2,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{56} = [(\sigma_s^2,0,0);(\sigma_s^2,0,0);(0,\sigma_s^2,\sigma_s^2);(0,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{57} = [(\sigma_s^2,0,0,0);(\sigma_{IR}^2+\sigma_s^2,\sigma_{IR}^2,0,0);(\sigma_{IR}^2,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2,\sigma_s^2);(0,\sigma_s^2,\sigma_s^2,\sigma_s^2)]$,
$\Sigma_{67} = [(\sigma_s^2,0,0,0);(0,\sigma_s^2,\sigma_s^2,\sigma_s^2);(0,\sigma_s^2,\sigma_{IR}^2+\sigma_s^2,\sigma_s^2)]$.

From the above, we deduce that

$\mathrm{Cov}_M(\mathbf{Y}_{I..},\mathbf{Y}_{II..}) = \mathbf{1}'\Sigma_{12}\mathbf{1}/9 = [\sigma_{IR}^2+4\sigma_s^2]/9$.
Similarly, we may deduce the rest of the covariance terms as follows:
$\mathrm{Cov}_M(\mathbf{Y}_{I..},\mathbf{Y}_{III..}) = \mathbf{1}'\Sigma_{13}\mathbf{1}/9 = [\sigma_{IR}^2+4\sigma_s^2]/9$,
$\mathrm{Cov}_M(\mathbf{Y}_{I..},\mathbf{Y}_{IV..}) = \mathbf{1}'\Sigma_{14}\mathbf{1}/12 = [\sigma_{IR}^2+5\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{I..},\mathbf{Y}_{V..}) = \mathbf{1}'\Sigma_{15}\mathbf{1}/12 = [\sigma_{IR}^2+6\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{I..},\mathbf{Y}_{VI..}) = \mathbf{1}'\Sigma_{16}\mathbf{1}/9 = [\sigma_{IR}^2+5\sigma_s^2]/9$,
$\mathrm{Cov}_M(\mathbf{Y}_{I..},\mathbf{Y}_{VII..}) = \mathbf{1}'\Sigma_{17}\mathbf{1}/12 = [\sigma_{IR}^2+7\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{II..},\mathbf{Y}_{III..}) = \mathbf{1}'\Sigma_{23}\mathbf{1}/9 = [\sigma_{IR}^2+5\sigma_s^2]/9$,
$\mathrm{Cov}_M(\mathbf{Y}_{II..},\mathbf{Y}_{IV..}) = \mathbf{1}'\Sigma_{24}\mathbf{1}/12 = [\sigma_{IR}^2+7\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{II..},\mathbf{Y}_{V..}) = \mathbf{1}'\Sigma_{25}\mathbf{1}/12 = [\sigma_{IR}^2+6\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{II..},\mathbf{Y}_{VI..}) = \mathbf{1}'\Sigma_{26}\mathbf{1}/9 = [\sigma_{IR}^2+4\sigma_s^2]/9$,
$\mathrm{Cov}_M(\mathbf{Y}_{II..},\mathbf{Y}_{VII..}) = \mathbf{1}'\Sigma_{27}\mathbf{1}/12 = [\sigma_{IR}^2+5\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{III..},\mathbf{Y}_{IV..}) = \mathbf{1}'\Sigma_{34}\mathbf{1}/12 = [\sigma_{IR}^2+7\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{III..},\mathbf{Y}_{V..}) = \mathbf{1}'\Sigma_{35}\mathbf{1}/12 = [\sigma_{IR}^2+6\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{III..},\mathbf{Y}_{VI..}) = \mathbf{1}'\Sigma_{36}\mathbf{1}/9 = [\sigma_{IR}^2+4\sigma_s^2]/9$,
$\mathrm{Cov}_M(\mathbf{Y}_{III..},\mathbf{Y}_{VII..}) = \mathbf{1}'\Sigma_{37}\mathbf{1}/12 = [\sigma_{IR}^2+5\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{IV..},\mathbf{Y}_{V..}) = \mathbf{1}'\Sigma_{45}\mathbf{1}/16 = [4\sigma_{IR}^2+8\sigma_s^2]/16$,
$\mathrm{Cov}_M(\mathbf{Y}_{IV..},\mathbf{Y}_{VI..}) = \mathbf{1}'\Sigma_{46}\mathbf{1}/12 = [\sigma_{IR}^2+5\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{IV..},\mathbf{Y}_{VII..}) = \mathbf{1}'\Sigma_{47}\mathbf{1}/16 = [4\sigma_{IR}^2+6\sigma_s^2]/16$,
$\mathrm{Cov}_M(\mathbf{Y}_{V..},\mathbf{Y}_{VI..}) = \mathbf{1}'\Sigma_{56}\mathbf{1}/12 = [\sigma_{IR}^2+6\sigma_s^2]/12$,
$\mathrm{Cov}_M(\mathbf{Y}_{V..},\mathbf{Y}_{VII..}) = \mathbf{1}'\Sigma_{57}\mathbf{1}/16 = [4\sigma_{IR}^2+8\sigma_s^2]/16$,
$\mathrm{Cov}_M(\mathbf{Y}_{VI..},\mathbf{Y}_{VII..}) = \mathbf{1}'\Sigma_{67}\mathbf{1}/12 = [\sigma_{IR}^2+7\sigma_s^2]/12$.

3. Data Analysis under Blinded Submission

We will now discuss essential features of data analysis for unbiased estimation of the finite population total $T(Y)$ of the study variable Y under the above interactive linear model. We refer to Hedayat and Sinha (1991) for standard results in finite population inference.

In a very general set-up, we have a finite labeled population of N units and we have taken recourse to a fixed size (n) sampling design with positive inclusion probabilities and joint inclusion probabilities for all pairs of units. For example, the well-known sampling design $SRSWOR(N,n)$ could be utilized.

Because of possible investigator and/or supervisor interventions, the response on the study variable Y_i for the selected RC(i) may be distorted and we stipulate a model as given above. For each sampling unit i in the sample, we simply take the average of the observations underlying it. Under the model assumptions, this serves as an unbiased estimate of the true response TR_i of the ith sample unit. We have used the notation $\mathrm{E}_M \mathbf{Y}_{i..} = TR_i$. Once this is ensured, we use the conventional Horvitz-Thompson Estimator [HTE, for short] for unbiased estimation of the population total $T(TR)$. In other words, we use $T(\hat{TR}) = \sum_i [(\hat{TR})_i]/\pi_i$, where $(\hat{TR})_i = \mathbf{Y}_{i..}$. An expression for Variance of $T(\hat{TR})$ has to be evaluated next. We use the standard formula: $V = V_1 E_2 + E_1 V_2$. Here E_2 and V_2 refer to model expectation and model variance. Clearly, model expectation results in the true values TR's. And then V_1 refers to computation of the variance of the HTE in terms of the TR's which is very much a standard exercise. Vide Hedayat and Sinha (1991) for details. For a fixed size (n) sampling design, this is given by

$$V_1 E_2 = \sum\sum_{i<j} [TR_i/\pi_i - TR_j/\pi_j]^2 (\pi_i \pi_j - \pi_{ij}).$$

Next, V_2 refers to the computation of model variance of the estimator based on the average responses for the sampled units. The estimator is the HTE for which the model variance involves all individual variances and all pair-wise covariances of the averages for the n sampled units. More explicitly,

$$\mathrm{V}_M(\sum_i \mathbf{Y}_{i..}/\pi_i) = \sum_i \mathrm{V}_M(\mathbf{Y}_{i..})/\pi_i^2 + \sum\sum_{i\neq j} \mathrm{Cov}_M(\mathbf{Y}_{i..}, \mathbf{Y}_{j..})/\pi_i \pi_j$$

All the entries involved in the above expression have already been worked out.

We will now discuss about the computation of $E_1 V_2$. Note that E_1 refers to expectation wrt the fixed size (n) sampling design. Here we need to go carefully. To discuss the general framework of such computations, we assume that $N = Mf$ so that all population units are grouped into M RCs of size f each. And we also set $m = nf$ as the total number of ultimate responding units so that in effect, our sampling design corresponds to a fixed size (n) sampling design for selection of n RCs each of size f, out of M clusters in the population, each of size f. The RCs are to be regarded as sampling units in our study and the cluster totals are the 'primary data' accrued from each sampled unit. These have been denoted by $Y_{[i;(j,k)]}$ for every triplet $[i;(j,k)]$ for which $I[i;(j,k)] = 1$.

Whenever investigator and / or supervisor interventions are likely to be present and are to be accounted for, we introduce, as in the above, an allocation matrix of order $t \times n$ to suggest the nature of allocation of the investigators among the sampled RCs and also another allocation matrix of order $q \times t$ to suggest the nature of supervisor-investigator 'dual checks' on the RCs' profiles. Here t denotes the number of investigators and q stands for the number of supervisors. While suggesting these two types of allocation matrices, we may take recourse to some 'nice' combinatorial structures such as BIBDs. The important point to be noted is that field data on each RC profile

should be collected/checked by more than one investigator and/or by more than one supervisor. In the illustrative example with $n = 7$, we chose $t = 7$ and $q = 3$ and the two matrices of orders $t \times n$ and $q \times t$ were presented in the same place through detailed descriptions of (RCI) to $RC(VII)$.

Having discussed these 'design issues', we are now in a position to project the concept underlying E_1. We treat the 'investigator-cum-RC' matrix of order $t \times n$ as 'given' once for all. So is the other matrix as well. Once the RCs are chosen according to a given fixed size (n) sampling design, we check the allocation matrix and adhere to it by assigning the columns of the allocation matrix to the sample RCs in ascending order of their labels. Since the two allocation matrices are chosen in advance, the variance and covariance computations underlying the interactive model will remain the same for all choices of the n RCs in the sample, except for their identification in terms of the RC-labels.

Thus, in the example above, we may decide on $N = 700$, $M = 70$, $f = 10$, $n = 7$, $t = 7$, $q = 2$, so that altogether 70 respondents are selected in 7 RCs of 10 each out of 70 RCs of 700 ultimate units. If the randomly selected RCs are labeled $[3, 17, 33, 41, 57, 63, 69]$, then the above expressions for model-based variance-covariances correspond to these RCs in the order mentioned. In other words, Σ_{11} in effect corresponds to Σ_{33} and so on. The actual realized RC-labels in ascending order take the positions of the labels $1, 2, \ldots, 7$ in the table and in the related computations.

It is evident that the exact computation of E_1 is quite involved. However, an unbiased sample RC-based estimator of E_1V_2 is simply given by V_2, once we assume the three variance components to be known.

Again, to find an unbiased estimator of V_1E_2, a trivial situation would have produced $\sum\sum_{i\neq j}[TR_i/\pi_i - TR_j/\pi_j]^2[\pi_i\pi_j - \pi_{ij}]/\pi_{ij}$ had the TR_i's been known. This follows from the standard result on variance estimation for the HTE viz., Yates–Grundy formula, as applicable for a fixed size (n) sampling design. Vide Hedayat and Sinha (1991), for example. However, in the present situation, TR_i's are unknown and instead we have the unbiased estimates of the TR_i's viz., $\mathbf{Y}_{i..} = \hat{TR_i}$. Therefore, we start with the expression $\sum\sum_{i\neq j}[\mathbf{Y}_{i..}/\pi_i - \mathbf{Y}_{i..}/\pi_j]^2[\pi_i\pi_j - \pi_{ij}]/\pi_{ij}]$ and work out its expectation i.e., E_1E_2. It follows that

$$E_2[...] = \sum\sum_{i\neq j}[TR_i/\pi_i - TR_j/\pi_j]^2[\pi_i\pi_j - \pi_{ij}]/\pi_{ij}] + \sum\sum_{i\neq j} \mathrm{V}_M[\mathbf{Y}_{i..}/\pi_i - \mathbf{Y}_{i..}/\pi_j][\pi_i\pi_j - \pi_{ij}]/\pi_{ij}].$$

Once more, if we assume the variance components to be known, then the second term above can be computed. Hence the first term above can be evaluated by subtraction.

Thus finally, we are in a position to derive an expression for the variance estimate, under the assumption of known variance components. The case of unknown variance components will be taken up in a subsequent communication.

Remark 3.1. At this stage, it may be mentioned that in a very general sense, we can make use of a BIBD(b, v, r, k, λ) for allocation of the $b = t$ investigators among the $v = n$ RCs with obvious interpretation of the other parameters $r,\ k,\ \lambda$. Note also that the investigator-supervisor allocation matrix does not necessarily have any structure or pattern, except that at least two supervisors need to be appointed and at least two should sit on each RC's field-based data file, which is already collected by independent venture of at least two investigators and made available for further scrutiny.

Remark 3.2. It must be noted that two distinct random processes are involved in the data analysis stage. One corresponds to the sampling design which results in the use of design-based unbiased estimator such as Horvitz-Thompson Estimator [HTE]. We need to work out variance estimate

corresponding to the HTE. On the other hand, the interactive linear model introduces model-based variance components which are assumed to be known. Therefore, variance estimation in this study refers to the sampling-design-based variance estimation from the survey data. Vide Hedayat and Sinha (1991) for details.

4. Illustrative Example [continued]

Without any loss of generality, we take the sample clusters to possess the labels $(1, 2, \ldots, 7)$. Further, we assume that the sampling design is $SRSWOR(M = 70,\ n = 7)$. The true population total is $T(TR)$ where each TR is composed of the sum of TR-values of 10 basic eus from within each of the clusters. We further assume that the reported data correspond to the subtotals based on within-cluster units. The true population subtotals are denoted by $TR_1,\ TR_2, \ldots,\ TR_{70}$ and our sample of size $n = 7$ provides model-based unbiased estimates for $TR_1,\ TR_2, \ldots,\ TR_7$. Further, since we adopt SRSWOR, we assert that

(i) unbiased estimate of $T(TR)$ is given by $M\times$ the sample average of within cluster estimates i.e., $T(\hat{TR}) = 10\sum_i \mathbf{Y}_{i..}$.

(ii) unbiased variance estimate is to be computed from

  (a) E_1V_2 component: It is just V_2 given by $M^2/n^2[\sum_i \mathrm{V}_M(\mathbf{Y}_{i..}) + \sum\sum_{i\neq j} \mathrm{Cov}_M(\mathbf{Y}_{i..}, \mathbf{Y}_{i..})]$

  (b) V_1E_2 component: It is the difference between two expressions given by
  First Expression: $[M^2(1/n - 1/M)][\sum\sum_{i<j}(\mathbf{Y}_{i..} - \mathbf{Y}_{j..})^2/n(n-1)]$;
  Second Expression: $[M^2(1/n - 1/M)][(n-1)\sum_i \sigma_{ii} - \sum\sum_{i\neq j}\sigma_{ij}]/n(n-1)$.

It follows, upon simplification, that

$$\sum_i \sigma_{ii} = 25/12\sigma_e^2 + 59/24\sigma_{IR}^2 + 143/36\sigma_s^2;$$
$$\sum\sum_{i<j} \sigma_{ij} = 22/9\sigma_{IR}^2 + 191/36\sigma_s^2.$$

By combining the two from (a) and (b) above, we obtain the final expression for the unbiased variance estimate as

$[M^2(1/n - 1/M)][\sum\sum_{i<j}(\mathbf{Y}_{i..} - \mathbf{Y}_{j..})^2/n(n-1)]$ [contribution from data]
PLUS
$[M/n][\sum_i \sigma_{ii}] + [M(M-1)/n(n-1)][\sum\sum_{i\neq j}\sigma_{ij}]$.

This latter expression simplifies to

$$[M/n][25/12\sigma_e^2 + 59/24\sigma_{IR}^2 + 143/36\sigma_s^2] + [2M(M-1)/n(n-1)][22/9\sigma_{IR}^2 + 191/36\sigma_s^2].$$

5. Interactive Linear Model under Unblinded Submission

Recall that $I[i;(j,k)] = 1$ if (j,k)-combination of the investigator and the supervisor have both worked on a schedule assigned to the ith responding unit. Therefore, if for a pair of triplets $I[i;(j,k)] = I[i;(j',k)] = 1$, under unblinded submission, supervisor labelled k has to handle two separate response profiles of the same respondent i and therefore, it only makes sense to first average out these two responses and then provide his/her own 'input' to that average before finalization of the response! Under blind submission, supervisor's input was incorporated for each response

profile submitted to him/her. That is the essential difference between the two scenarios.
For the same example as before, we now sort out the final scheme of 'averaging' as follows:

(I) $Y_{[I^*]} = [(Y_{I,(1,1)}) + 1/2[Y_{I,(5,2)} + Y_{I,(7,2)}]]/2 = [2Y_{I,(1,1)} + Y_{I,(5,2)} + Y_{I,(7,2)}]/4;$
(II) $Y_{[II^*]} = [(Y_{II,(1,1)} + Y_{II,(2,1)})/2 + Y_{II,(6,2)}]/2 = [Y_{II,(1,1)} + Y_{II,(2,1)} + 2Y_{II,(6,2)}]/4;$
(III) $Y_{[III^*]} = [(Y_{III,(2,1)} + Y_{III,(3,1)})/2 + Y_{III,(7,2)}]/2 = [Y_{III,(2,1)} + Y_{III,(3,1)} + 2Y_{III,(7,2)}]/4;$
(IV) $Y_{[IV^*]} = [(Y_{IV,(1,1)} + Y_{IV,(3,1)} + Y_{IV,(4,1)})/3 + Y_{IV,(4,2)}]/2$
$= [Y_{IV,(1,1)} + Y_{IV,(3,1)} + Y_{IV,(4,1)} + 3Y_{IV,(4,2)}]/6;$
(V) $Y_{[V^*]} = [(Y_{V,(2,1)} + Y_{V,(4,1)}) + (Y_{V,(4,2)} + Y_{V,(5,2)})]/4;$
(VI) $Y_{[VI^*]} = [(Y_{VI,(3,1)}) + 1/2[Y_{VI,(5,2)} + Y_{VI,(6,2)}]]/2 = [2Y_{VI,(3,1)} + Y_{VI,(5,2)} + Y_{VI,(6,2)}]/4;$
(VII) $Y_{[VII^*]} = [(Y_{VII,(4,1)}) + 1/3[Y_{VII,(4,2)} + Y_{VII,(6,2)} + Y_{VII,(7,2)}]]/2$
$= [3Y_{VII,(4,1)} + Y_{VII,(4,2)} + Y_{VII,(6,2)} + Y_{VII,(7,2)}]/6.$

Next, we need to compute variances and covariances of the resulting input averages from all the seven respondent groups. These are computed below.
Recall Σ_{11}, representing the dispersion matrix of $Y_{[I;(1,1)]}; Y_{I;(5,2)]}; Y_{[I;(7,2)]}$, has already been derived in the form $[(\sigma_e^2 + \sigma_{IR}^2 + \sigma_S^2, 0, 0); (0, \sigma_e^2 + \sigma_{IR}^2 + \sigma_S^2, \sigma_S^2); (0, \sigma_S^2, \sigma_e^2 + \sigma_{IR}^2 + \sigma_S^2)]$.
Therefore, $\mathrm{V}_M\mathbf{Y}_{[I^*]} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 4\sigma_S^2]/8$, upon simplification.
Likewise, for the other variance components, the expressions are given below.

$\mathrm{V}_M\mathbf{Y}_{[II^*]} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 4\sigma_S^2]/8$, upon simplification.
$\mathrm{V}_M\mathbf{Y}_{[III^*]} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 4\sigma_S^2]/8$, upon simplification.
$\mathrm{V}_M\mathbf{Y}_{[IV^*]} = [2\sigma_e^2 + 3\sigma_{IR}^2 + 5\sigma_S^2]/8$, upon simplification.
$\mathrm{V}_M\mathbf{Y}_{[V^*]} = [2\sigma_e^2 + 3\sigma_{IR}^2 + 4\sigma_S^2]/8$, upon simplification.
$\mathrm{V}_M\mathbf{Y}_{[VI^*]} = [3\sigma_e^2 + 3\sigma_{IR}^2 + 4\sigma_S^2]/8$, upon simplification.
$\mathrm{V}_M\mathbf{Y}_{[VII^*]} = [2\sigma_e^2 + 3\sigma_{IR}^2 + 3\sigma_S^2]/6$, upon simplification.

Next, towards computation of the covariance terms, we note that Σ_{ij} matrices are already displayed before. Therefore, we only do the computations of the form $\mathbf{x}'\Sigma\mathbf{y}$ for different choices of $\mathbf{x}$ and $\mathbf{y}$ as are relevant for the averages.

(i) $\mathrm{Cov}_M(\mathbf{Y}_{[I^*]}, \mathbf{Y}_{[II^*]}) = (2\ 1\ 1)'\Sigma_{12}(1\ 1\ 2)/16 = [2\sigma_{IR}^2 + 8\sigma_s^2]/16;$
(ii) $\mathrm{Cov}_M(\mathbf{Y}_{[I^*]}, \mathbf{Y}_{[III^*]}) = (2\ 1\ 1)'\Sigma_{13}(1\ 1\ 2)/16 = [2\sigma_{IR}^2 + 8\sigma_s^2]/16;$
(iii) $\mathrm{Cov}_M(\mathbf{Y}_{[I^*]}, \mathbf{Y}_{[IV^*]}) = (2\ 1\ 1)'\Sigma_{14}(1\ 1\ 2\ 2)/24 = [2\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(iv) $\mathrm{Cov}_M(\mathbf{Y}_{[I^*]}, \mathbf{Y}_{[V^*]}) = (2\ 1\ 1)'\Sigma_{15}(1\ 1\ 2\ 2)/24 = [2\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(v) $\mathrm{Cov}_M(\mathbf{Y}_{[I^*]}, \mathbf{Y}_{[VI^*]}) = (2\ 1\ 1)'\Sigma_{16}(2\ 1\ 1)/16 = [1\sigma_{IR}^2 + 8\sigma_s^2]/16;$
(vi) $\mathrm{Cov}_M(\mathbf{Y}_{[I^*]}, \mathbf{Y}_{[VII^*]}) = (2\ 1\ 1)'\Sigma_{17}(3\ 1\ 1\ 1)/24 = [1\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(vii) $\mathrm{Cov}_M(\mathbf{Y}_{[II^*]}, \mathbf{Y}_{[III^*]}) = (1\ 1\ 2)'\Sigma_{23}(1\ 1\ 2)/16 = [\sigma_{IR}^2 + 8\sigma_s^2]/16;$
(viii) $\mathrm{Cov}_M(\mathbf{Y}_{[II^*]}, \mathbf{Y}_{[IV^*]}) = (1\ 1\ 2)'\Sigma_{24}(1\ 1\ 2\ 2)/24 = [\sigma_{IR}^2 + 14\sigma_s^2]/24;$
(ix) $\mathrm{Cov}_M(\mathbf{Y}_{[II^*]}, \mathbf{Y}_{[V^*]}) = (1\ 1\ 2)'\Sigma_{25}(1\ 1\ 2\ 2)/24 = [\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(x) $\mathrm{Cov}_M(\mathbf{Y}_{[II^*]}, \mathbf{Y}_{[VI^*]}) = (1\ 1\ 2)'\Sigma_{26}(2\ 1\ 1)/16 = [2\sigma_{IR}^2 + 8\sigma_s^2]/16;$
(xi) $\mathrm{Cov}_M(\mathbf{Y}_{[II^*]}, \mathbf{Y}_{[VII^*]}) = (1\ 1\ 2)'\Sigma_{27}(3\ 1\ 1\ 1)/24 = [2\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(xii) $\mathrm{Cov}_M(\mathbf{Y}_{[III^*]}, \mathbf{Y}_{[IV^*]}) = (1\ 1\ 2)'\Sigma_{34}(1\ 1\ 2\ 2)/24 = [\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(xiii) $\mathrm{Cov}_M(\mathbf{Y}_{[III^*]}, \mathbf{Y}_{[V^*]}) = (1\ 1\ 2)'\Sigma_{35}(1\ 1\ 2\ 2)/24 = [\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(xiv) $\mathrm{Cov}_M(\mathbf{Y}_{[III^*]}, \mathbf{Y}_{[VI^*]}) = (1\ 1\ 2)'\Sigma_{36}(2\ 1\ 1)/16 = [2\sigma_{IR}^2 + 8\sigma_s^2]/16;$
(xv) $\mathrm{Cov}_M(\mathbf{Y}_{[III^*]}, \mathbf{Y}_{[VII^*]}) = (1\ 1\ 2)'\Sigma_{373}\ 1\ 1\ 1)/24 = [2\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(xvi) $\mathrm{Cov}_M(\mathbf{Y}_{[IV^*]}, \mathbf{Y}_{[V^*]}) = (1\ 1\ 2\ 2)'\Sigma_{45}(1\ 1\ 2\ 2)/36 = [12\sigma_{IR}^2 + 16\sigma_s^2]/36;$
(xvii) $\mathrm{Cov}_M(\mathbf{Y}_{[IV^*]}, \mathbf{Y}_{[VI^*]}) = (1\ 1\ 2\ 2)'\Sigma_{46}(2\ 1\ 1)/24 = [2\sigma_{IR}^2 + 10\sigma_s^2]/24;$

(xviii) $\text{Cov}_M(\mathbf{Y}_{[IV^*]}, \mathbf{Y}_{[VII^*]}) = (1\ 1\ 2\ 2)'\Sigma_{47}3\ 1\ 1\ 1)/36 = [16\sigma_{IR}^2 + 18\sigma_s^2]/16;$
(xix) $\text{Cov}_M(\mathbf{Y}_{[V^*]}, \mathbf{Y}_{[VI^*]}) = (1\ 1\ 2\ 2)'\Sigma_{56}(2\ 1\ 1)/24 = [2\sigma_{IR}^2 + 12\sigma_s^2]/24;$
(xx) $\text{Cov}_M(\mathbf{Y}_{[V^*]}, \mathbf{Y}_{[VII^*]}) = (1\ 1\ 2\ 2)'\Sigma_{57}3\ 1\ 1\ 1)/36 = [12\sigma_{IR}^2 + 18\sigma_s^2]/36;$
(xxi) $\text{Cov}_M(\mathbf{Y}_{[VI^*]}, \mathbf{Y}_{[VII^*]}) = (2\ 1\ 1)'\Sigma_{67}(3\ 1\ 1\ 1)/24 = [1\sigma_{IR}^2 + 12\sigma_s^2]/24.$

In the next section, we discuss about relevant changes in the data analysis.

6. Data Analysis under Unblinded Submission

We will closely follow the data analysis in the blinded submission case and only suggest the changes relevant to the current scenario.

(i)* unbiased estimate of $T(TR)$ is given by $M\times$ the sample average of within cluster estimates i.e., $T(\hat{TR}) = 10[\mathbf{Y}_{[I*]} + \cdots + \mathbf{Y}_{[VII*]}]$.
(ii)* unbiased variance estimate is to be computed from
  (a)* E_1V_2 component: It is just V_2 given by $M^2/n^2[\sum_i \text{V}_M(\mathbf{Y}_{[I*]}) + \sum\sum_{i\neq j}\text{Cov}_M(\mathbf{Y}_{[I*]}, \mathbf{Y}_{[J*]})]$
  (b)* V_1E_2 component: It is the difference between two expressions given by
  First Expression: $[M^2(1/n - 1/M)][\sum\sum_{i<j}(\mathbf{Y}_{[I*]} - \mathbf{Y}_{[J*]})^2/n(n-1)]$;
  Second Expression: $[M^2(1/n - 1/M)][(n-1)\sum_i \sigma_{ii}^* - \sum\sum_{i\neq j}\sigma_{ij}^*]/n(n-1)$.

In the above, σ_{ii}^*'s refer to variances of $Y_{[I*]}$'s and σ_{ij}^*'s refer to the covariances of $Y_{[I*]}, Y_{[II*]}$'s.

It follows, upon simplification, that

$$\sum_i \sigma_{ii}^* = 7/3\sigma_e^2 + 11/4\sigma_{IR}^2 + 7/2\sigma_s^2;$$
$$\sum\sum_{i<j}\sigma_{ij}^* = 179/72\sigma_{IR}^2 + 94/9\sigma_s^2.$$

By combining the two from (a)* and (b)* above, we obtain the final expression for the unbiased variance estimate as

$[M^2(1/n - 1/M)][\sum\sum_{i<j}(\mathbf{Y}_{[I*]} - \mathbf{Y}_{[II*]})^2/n(n-1)]$ [contribution from data]
PLUS
$[M/n][\sum_i \sigma_{ii}^*] + [M(M-1)/n(n-1)][\sum\sum_{i\neq j}\sigma_{ij}^*]$.

This latter expression simplifies to

$$[M/n][7/3\sigma_e^2 + 11/4\sigma_{IR}^2 + 7/2\sigma_s^2] + [2M(M-1)/n(n-1)][179/72\sigma_{IR}^2 + 94/9\sigma_s^2].$$

7. Concluding Observations

We believe that this is a modest start of a long-going project on study of interactive linear models in survey sampling context, depicting the simplest-to-intricate involvements of the investigators and/or supervisors in the data-gathering process till it reaches the data collection agencies. We have assumed the variance components to be known which is a serious limitation of the study. The full potential of mixed linear models has yet to be explored for estimation of the variance components.

Acknowledgement

We are indeed thankful to two anonymous referees for pointing out several mistakes/typos and for their insightful comments and observations which have been very helpful in carrying out the revision of the submitted version.

References

[1] Hedayat, A.S. and Sinha, Bikas K. (1991). *Design and Inference in Finite Population Sampling*. Wiley, New York.
[2] Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. Wiley, New York.
[3] Searle, S.R. (1971). *Linear Models*. Wiley, New York.