Overview of the Simulation Study for Performance of Balanced Two Stage Predictors of Realized Random Cluster Means

Ed Stanek

Introduction

We plan a simulation study similar to the study reported by Pfeffermann and Nathan (1981, JASA 681-689). The simulation enables evaluation of predictors in two stage cluster sampling contexts such as the Seasons study. In that study, there were 3 measures on each cluster (subject) in each 3-month period. We assume that we are interested in estimating the realized cluster mean. We may be interested in a cluster mean over different time periods, such as a 7 day average, a 30 day average, and a 365 day average. An earlier description of the simulation is contained in c03ed10.doc.

The Population

We define the population via a set of parameters generated using random numbers prior to the simulation. Initially, we will use normal distribution random number generators. We may later define a population using other schemes, such as empirical data from the Seasons study, and model responses around a substantive variable. We describe the process for developing the population generated by normal random numbers next.

For all simulations, the population mean μ will be set to a constant value. The population will be composed of N clusters, with M units per cluster. We will generate a cluster mean by defining an initial variance between cluster parameters, σ_P^2. Using this variance, we will generate N random variables using a $N\left(\mu,\sigma_P^2\right)$ distribution. These values will form the basis for the cluster parameters. Since the number of clusters in the population is finite, the average of the cluster parameters will not necessarily equal the population mean, μ. We will force the mean of the cluster parameters to equal μ by centering the cluster parameters at μ. Each individual cluster parameters will be represented as μ_s. We round the values of these parameters to the nearest hundredth to simplify presentation of simulation results. Finally, the variance of the cluster parameters in the generated population will not equal σ_P^2 since, once again, the number of clusters is finite. We will define the variance of the cluster parameters as $\sigma^2 = \sum_{s=1}^{N} \frac{\left(\mu_s - \mu\right)^2}{N-1}$ using the simulated cluster means.

Next, we will generate unit effects for the M units for each cluster. We will force these effects to average to zero. We generate the unit effects using a normal random variable generator, $N\left(0,\sigma_{Ps}^2\right)$. Since the sum of the unit effects for the finite population will not be exactly zero, we will center these effects so that their sum is exactly zero. The parameter for a cluster-unit will be formed by adding the unit effect to the cluster mean, and represented by μ_{st}. The variance of the unit parameters for a cluster will not equal σ_{Ps}^2 since once again, the number of units per cluster is finite. We

will define the variance of the unit parameters as $\sigma_s^2 = \sum_{t=1}^{M} \frac{(\mu_{st} - \mu_s)^2}{M-1}$ using the generated unit parameters. Notice that we will only generate one set of unit effects for the population. These unit effects will be used to create parameters for cluster-units for each cluster. We generate the unit effects in this manner to force the variance of units within clusters to be constant for all clusters. The common within cluster variance which we represent by σ_e^2 is equal to the average within cluster variance, $\sigma_e^2 = \sum_{s=1}^{N} \frac{\sigma_s^2}{N}$.

Overview of the Simulation of the Population Parameters.

The simulation program will consist of modules. The first module generates parameters for the finite population. Response error will not be included in this generation. The population parameters are stored in a data file with one record for each cluster. Unit parameters correspond to columns in the data file.

Table 1a. Variable Definitions in the Simulation for the Population of Clusters

Variable	**SAS Name**	**Description**
N	&bn	# of clusters in the population
M	&bm	# of units for a cluster
μ	&mu_p	Population average
σ_P^2	&popv	Initial variance of cluster means.
σ_{Ps}^2	&popcv	Initial variance of units in a cluster.
μ_s	mu_s	Cluster parameter
y_{st}	t1-tM	Cluster-unit parameters
σ^2	v_mus	Variance of cluster parameters (over N-1)
σ_s^2	v_s	Variance of units for a cluster (over M-1)
$\sigma_e^2 = \sum_{s=1}^{N} \frac{\sigma_s^2}{N}$	v_e	Average variance of units for a cluster (over M-1)
s	s	Cluster label

Table 1b. Variables Used to Simulate the Population, and then Dropped

Variable	SAS Name	Description
outpx	&outpx	Variable to control descriptive output about simulated population (0=none, 1=minimal, 2= detailed)
$\sum_{s=1}^{N} \mu_s^*$	ssum	Sum of initial generated cluster means (that does not equal the population average
μ_s^*	ys1-ysN	Initial generated cluster means
$\frac{\sum_{s=1}^{N} \mu_s^*}{N} - \mu$	scr_mu	Correction to make pop. mean equal to designed mean
σ^2	v_musx	Temporary name for variance of cluster parameters.
	match	Value of 1 used to combine data sets.
$\sum_{t=1}^{M} \left(\mu_{st}^* - \mu_s \right)$	ssumc	Sum of initial generated unit deviations for cluster (that do not sum to zero)
$\mu_{st}^* - \mu_s$	t1-tM	Initial generated unit value deviations for a cluster.
$\frac{\sum_{t=1}^{M} \left(\mu_{st}^* - \mu_s \right)}{M}$	scr_muc	Correction to make unit deviations sum to zero in clusters.
$\sum_{t=1}^{M} \left(\mu_{st} - \mu_s \right)^2$	sqsum	Sums of squared unit deviations for unit variance

Selection of Two Stage Cluster Samples

Selection of a Sample of Clusters

The second module in the simulation will select a two stage simple random sample from a clustered population. First, using a list of cluster labels, a simple random sample of cluster labels is identified. This process is repeated many times (with each repetition called a 'trial'). Once the sampled clusters are identified, they are combined with the population data, and from these data, a sample of units in each cluster is selected. Input to the module is a finite list of fixed values (variables) corresponding to the values for units in the population.

We describe construction of the basic permutation in more detail. As an example, suppose that there are N clusters in the list to be permuted. Let each cluster be associated with a label s. Also, let us set the value of each cluster to be equal to the cluster label. To form a permutation, we first select one of the clusters. For the first position in the permutation, any one of the $N^* = N$ clusters will have the same chance of

being selected, i.e. $\frac{1}{N^*}$. We make the selection by dividing the 0-1 interval into N^* equal size intervals. The starting and ending point for interval p is given by $\frac{p-1}{N^*}$ and $\frac{p}{N^*}$. A uniform random number generator is used to select a number between 0 and 1, and the result is used to identify the cluster that is picked in the selection. We place the parameter for this cluster in position $i = 1$ in a list of permuted values.

After selection of the first cluster, there will be $N^* = N - 1$ remaining clusters in the population. Prior to selecting the next cluster in the permutation, we re-assign the labels to the clusters remaining in the population. For each cluster in the list with a label greater than the selected cluster, we reduce the label by one. For example, suppose that there are $N = 5$ clusters in the population, and that the cluster with label $s = 3$ was selected as the first cluster in a permutation. Then the remaining $N^* = 4$ clusters would be re-assigned such that cluster values appeared as the values 1, 2, 4, 5.

We select the second cluster in a permutation by dividing the 0-1 interval into N^* equal size intervals. For the second selection, $N^* = N - 1$. The starting and ending point for interval p is given by $\frac{p-1}{N^*}$ and $\frac{p}{N^*}$. A uniform random number generator is used to select a number between 0 and 1, and the result is used to identify the cluster that is picked in the selection. We place the parameter for this cluster in position $i = 2$ in a list of permuted values. Continuing in this process, we may form a permutation of all the values in the population. Alternatively, we may stop after generating the permuted values for $i = 1,...,n$ positions. In the simulation program, we only generate the first n positions for each permutation, since these positions define the sample.

Once a set of clusters is selected for a sample, we associate with each selected cluster the unit values for the cluster. A similar process is then followed for selecting a simple random sample of unit values from the selected clusters. The results of these selections are written to a data set for a two-stage cluster sample (referred to as a trial). This process is repeated for many trials.

The first part of the selection macro selects a sample of clusters. One data set is created for reference (C1, with 1 record per trial), while a second data set with one record per selected cluster per trial (C2) is combined with the population data (P3) to form the data set SAMP1 prior to selecting units in a cluster.

Table 2a. Variables generated in Simulating a Sample of Clusters Kept in a Data set (C1) with One record per trial.

Variable	SAS Name	Description
Trials	&trialn	Number of independent samples selected from the population.
N	&bn	# of clusters in the population
M	&bm	# of units for a cluster
ns	&ns	Number of clusters selected in a sample.
ms	&ms	Number of units selected in a selected cluster.
$\bar{\bar{\sigma}}^2$	&vbb	Population Average of response error variance
chk	&chk	# of replications of permutations of units (for checking)
outpxc	&outpxc	Variable to control descriptive output about simulated sample clusters (0=none, 1=minimal, 2= detailed)
outpxu	&outpxu	Variable to control descriptive output about simulated sample units in a cluster (0=none, 1=minimal, 2= detailed)
trial	trial	The trial number.
si1-sin	si1-sin	An array of cluster values that are shifted after each selection of a cluster in the permutation.

Table 2b. Variables generated in Simulating a Sample of Clusters Kept in a Data set (C2) with One record per position in a permutation per trial.

Variable	SAS Name	Description
Trials	&trialn	Number of independent samples selected from the population.
ns	&ns	Number of clusters selected in a sample.
trial	trial	The trial number.
s	s	Label for selected cluster in a permutation.
i	i	The position in the permutation of the selected cluster.

Table 2c. Variables Used to Permute Clusters in the Population, and then Dropped from the Two-Stage Permutation Macro

Variable	SAS Name	Description
s	u	Temporary label for cluster in the population that is assigned as the value for a cluster for permuting clusters.
N^*	rem_p	The number of remaining clusters to permute in the population.
p	p	The relative position of the clusters remaining to be permuted in the population. If there N clusters, then p initially p=1,…,N. After selecting 2 clusters, p=1,…,N-2.
rem_s	rem_s	The number of clusters remaining to select in a sample.
rn	rn	The random number selected from the 0-1 generator
select	select	An indicator having a value of 0 if clusters in all previous positions have not been chosen for a particular selection, or a value of 1 if a cluster in the current or previous position has been selected.
sx1-sxN	sx1-sxN	An array of cluster values set equal to the labels 1 to N.

Selection of a Sample of Units in a Cluster

The second stage sample is selected from the selected clusters in the data set SAMP1. The process of selection is similar to the selection of clusters. The resulting sample is stored in the data set SAMP2. This data set is the simulated two stage sample data.

Table 3a. Variables generated in Simulating a Sample of Units for a sample of Clusters Kept in a Data set (SAMP2) with one record selected cluster per trial.

Variable	**SAS Name**	**Description**
Trials	&trialn	Number of independent samples selected from the population.
s	s	Label for the cluster
N	&bn	# of clusters in the population
M	&bm	# of units for a cluster
ms	&ms	Number of units selected in a selected cluster.
μ	&mu_p	Population average
μ_s	mu_s	Cluster parameter
σ^2	v_mus	Variance of cluster parameters (over N-1)
σ_s^2	v_s	Variance of units for a cluster (over M-1)
$\bar{\bar{\sigma}}^2$	&vbb	Population Average of response error variance
trial	trial	The trial number
outpxu	&outpxu	Variable to control descriptive output about simulated sample clusters (0=none, 1=minimal, 2= detailed)
Y_{ij}	j1-jm	Sample value for units without response error
Y_{ijk}	jre1-jrem	Sample values for units with response error

Table 3b. Variables generated in Simulating a Sample of Units for a sample of Clusters Kept in a Data set (STEST) to Test permutations of units within clusters.

Variable	**SAS Name**	**Description**
outpxu	&outpxu	Variable to control descriptive output about simulated sample clusters (0=none, 1=minimal, 2= detailed)
trial	trial	The trial number
chk	&chk	# of times to permute units in the 1st selected cluster as a test.
tu	tu	Replication number for a permutation of units for a selected cluster.
	chkt1-chktM	Array used in checking permutations of units for positions. This is set equal to the value of the unit parameters for a selected cluster. It is needed so that multiple permutations of units in the same cluster begin with the same set of cluster units.

Table 3c. Variables generated in Simulating a Sample of Units for a sample of Clusters and dropped from created data sets

Variable	SAS Name	Description
y_{it}	t1-tM	Array of unit values for ith selected cluster
N^*	rem_pc	The number of remaining clusters to permute in the population.
p	unit	The relative position of the units remaining to be permuted in a selected cluster. If there M units/clusters, then unit initially unit=1,…,M. After selecting 2 units, unit=1,…,M-2.
rem_sc	rem_sc	The number of units remaining to select in a selected cluster.
rn	rn	The random number selected from the 0-1 generator
selectc	selectc	An indicator having a value of 0 if units in all previous positions have not been chosen for a particular selection, or a value of 1 if a unit in the current or previous position has been selected.

Response Error

We add response error that follows a normal distribution, $N\left(0,\bar{\bar{\sigma}}^2\right)$ to a cluster-unit parameter using a random number generator. The response error for all cluster-units is set equal for all units in all clusters.

Predictors in the Simulation

Predictors are constructed for each trial. To develop a predictor, we use quantities calculated for each selection of a cluster (PSU), and averages of these quantities over selected clusters (PSUs) in the sample (trial). The predictor is obtained by combining these results.

We describe this process based on development of the mixed model predictor (given by equation (1.4) in Stanek and Singer(2003)). First, we evaluate $\sigma_i^2 = \sum_{s=1}^{N} u_{is}\left(\sigma_s^2 + \frac{1}{m}\sum_{j=1}^{m}\sum_{t=1}^{M} u_{jt}^{(s)}\sigma_{st}^2\right)$. This corresponds to the realized variance for the i^{th} selected cluster. Then, for each selected cluster, we evaluate the quantities $\bar{\bar{Y}}_i$, and $v_i = \sigma^2 + \frac{\sigma_i^2}{m}$. We also form the quantities $\frac{\bar{\bar{Y}}_i}{v_i}$ and $\frac{1}{v_i}$.

The next step is to sum these quantities over the selected clusters in the trial resulting in $\sum_{i=1}^{n}\frac{\bar{\bar{Y}}_i}{v_i}$ and $\sum_{i=1}^{n}\frac{1}{v_i}$, from which we determine $\hat{\mu} = \sum_{i=1}^{n} w_i \bar{\bar{Y}}_i$ where $w_i = \frac{1/v_i}{\sum_{i^*=1}^{n} 1/v_{i^*}}$,

equivalent to $\hat{\mu} = \dfrac{\sum_{i=1}^{n} \dfrac{\overline{Y}_i}{v_i}}{\sum_{i=1}^{n} \dfrac{1}{v_i}}$. Finally, combining these results back into the sample, we obtain the mixed model predictor given by $\hat{p} = \hat{\mu} + k_i\left(\overline{Y}_i - \hat{\mu}\right)$.

We summarize the variables used to develop the predictors in Table 4a.

Table 4. Variable Definitions in the Simulation for Developing Predictors

Variable	**SAS Name**	**Description**
Arrays		
$Y_{i1} - Y_{im}$	j{m}	Response for ith selected cluster.
$Y_{11} - Y_{nm}$	sp{n,m}	Response for selected clusters and units
$\overline{Y}_1 - \overline{Y}_n$	yibx{n}	Average of selected unit responses for selected cluster
$\dfrac{1}{v_1} - \dfrac{1}{v_n}$	wib{n}	Inverse of variance of selected sample cluster mean
$\dfrac{\overline{Y}_1}{v_1} - \dfrac{\overline{Y}_n}{v_n}$	yibw{n}	Weighted average of selected unit responses for selected cluster
v_i	vib	# of clusters in the population
$\overline{\overline{Y}}$	ybb	Sample mean of clusters for trial
$\hat{\mu}$	muwhat	Weighted average sample mean
$\sigma^{*2} = \sigma^2 - \dfrac{\sigma_e^2}{M}$	vstar	Adj. Variance of selected cluster for RP Model
$f = \dfrac{m}{M}$	f	Unit in cluster sampling fraction
k_i	ki	Mixed Model shrinkage constant
k	k	RP Model shrinkage constant
k^*	kstar	RP Model shrinkage constant with denominator response error
k_r^*	krstar	RP Model shrinkage constant with numerator unit variance and denominator response error
σ_i^2	vi	Variance of selected cluster and response error
$\overline{Y}_i$	yib	Sample mean for selected cluster
a	a1-a5	Coefficients for a in the predictors $\hat{P} = a\overline{Y} + b\overline{Y}_i$

b	b1-b5	Coefficients for b in the predictors $\hat{P} = a\overline{Y} + b\overline{Y}_i$
p	p1-p5	Predictor p given by $\hat{P} = a\overline{Y} + b\overline{Y}_i$ corresponding to mean, mixed model, Scott and Smith, RP model, and RP model with response error predictors.
	tmse1-tmse5	Theoretical MSE for the predictor
	sqd1-sqd5	Squared deviation of predictor from true value
# Trials*n	ntot	Number of sampled clusters in simulation.
$\frac{\sum_{Trials}\left(\sum_{i=1}^{n}\left(\overline{Y}_i - \mu_i\right)^2\right)}{Trials}$	mse1	Simple Cluster Average-Average MSE
$\frac{\sum_{Trials}\left(\sum_{i=1}^{n}\left(\hat{p}_i - \mu_i\right)^2\right)}{Trials}$	mse2	Mixed Model Average MSE
$\frac{\sum_{Trials}\left(\sum_{i=1}^{n}\left(\hat{P}_i - \mu_i\right)^2\right)}{Trials}$	mse3	Scott and Smith's Average MSE
$\frac{\sum_{Trials}\left(\sum_{i=1}^{n}\left(\hat{T}_i - \mu_i\right)^2\right)}{Trials}$	mse4	RP Model Average MSE (no response error)
$\frac{\sum_{Trials}\left(\sum_{i=1}^{n}\left(\hat{T}_i - \mu_i\right)^2\right)}{Trials}$	mse5	RP Model Average MSE (with response error)

We summarize the predictors in the simulation. They are given as follows:

Table 1. Predictors of the Latent Value of PSU i when $i \le n$ in Two-stage Cluster Sampling

Model	**Predictor**	**SAS Name**
Mean	$\bar{Y}_i^*$	p1
Mixed Model	$\hat{p} = \left(\hat{\mu} + k_i\left(\bar{Y}_i^* - \hat{\mu}\right)\right)$	p2
Scott&Smith	$\hat{\bar{P}}_i = f\bar{Y}_i^* + (1-f)\left(\hat{\mu}^* + k_i^*\left(\bar{Y}_i^* - \hat{\mu}^*\right)\right)$	p3
Random. Perm.	$\hat{\bar{T}}_i = f\bar{Y}_i^* + (1-f)\left(\bar{Y}^* + k\left(\bar{Y}_i^* - \bar{Y}^*\right)\right)$	p4
RP + Resp. Err.	$\hat{\bar{T}}_i = f\left(\bar{Y}^* + k_r^*\left(\bar{Y}_i^* - \bar{Y}^*\right)\right) + (1-f)\left(\bar{Y}^* + k^*\left(\bar{Y}_i^* - \bar{Y}^*\right)\right)$	p5
Est. MM	$\hat{\hat{p}} = \left(\hat{\mu} + \hat{k}_i\left(\bar{Y}_i^* - \hat{\mu}\right)\right)$	p6
Est. Scott&Smith	$\hat{\hat{\bar{P}}}_i = f\bar{Y}_i^* + (1-f)\left(\hat{\mu}^* + \hat{k}_i^*\left(\bar{Y}_i^* - \hat{\mu}^*\right)\right)$	p7
Est. RP	$\hat{\hat{\bar{T}}}_i = f\bar{Y}_i^* + (1-f)\left(\bar{Y}^* + \hat{k}\left(\bar{Y}_i^* - \bar{Y}^*\right)\right)$	p8
Est RP + Resp. Err.	$\hat{\hat{\bar{T}}}_i = f\left(\bar{Y}^* + \hat{k}_r^*\left(\bar{Y}_i^* - \bar{Y}^*\right)\right) + (1-f)\left(\bar{Y}^* + \hat{k}^*\left(\bar{Y}_i^* - \bar{Y}^*\right)\right)$	p9

Mean squared errors are organized in a similar manner.

The Simulation

The simulation is developed in a set of program. The programs are given by

ced03p23.sas Separate macro to develop population and sampling (no predictors)
ced03p24.sas Separate macro for predictors reading in sample
ced03p25.sas Separate macros for population and sampling, with predictors
ced03p26.sas A single macro is used to generate the population, samples, and predictors.
ced03p29.sas Working version of simulation including response error.

The program ced03p29.sas using these parameters in the macro:

```
%simc(25,   /*# Clusters in population: N   */
      10,    /*# Units in a cluster:    M   */
      50,    /*Population mean          mu  */
       8,    /*Var Between Cluster means    */
      40,    /*Var Between Units in clusters*/
    1000,    /*# of trials                  */
      20,    /*# Clusters in sample : n     */
       2,    /*# Number of Units per cluster in sample: m*/
     160,    /*# Number of units in sample=n*m           */
       0,    /*Response error on cluster-unit*/
       1,    /*# samples per selected cluster. Set=1 unless you want to
                check the permutations*/
       0,    /*Value to control printed output for population:
                0=none, 1=limited, 2=a lot*/
      0,    /*Value to control printed cluster selection output:
               0=none, 1=limited, 2=a log*/
      0,    /*Value to control printed unit selection output:
               0=none, 1=limited, 2=a lot*/
      0     /*Value to control printed output for development of
              predictors: 0=none 1=limited, 2=a lot*/
       );
```

We include the results of several simulations as an illustration. Note that the RP model predictors have smaller MSE in all settings. Also note that the simple sample mean may do better than the Mixed Model BLUP in some cases. There is close correspondence between the theoretical MSE and the simulated MSE in all cases. The theoretical MSE is developed in c03ed11.doc. With response error, the theoretical MSE is developed in c03ed12.doc.

Examples of Simulation Output (Using ced03p26.sas).

Source:ced03p26.sas C:\projects\cluster\programs 5/8/03 EJS
Table 26-1. Summary Population Parameters

# Pop Cluster: BN	# Pop Units: BM	Pop Mean: MU_P	Cluster Var: V_MUS	Pop Ave Unit Var: V_E
25	10	50.0000	4.9820	20.7276

Table 26-2. Summary of MSE results: N=25 (n=20) M=10 (m=2) Pop-Mean=50 Resp.Error-Var=0

Total # of Simulated Clusters: NTOT	Statistic Parameter Differnce: TYP	Theory Cluster Mean: TMSE1	Theory MSE MM BLUP: TMSE2	Theory MSE SS BLUP: TMSE3	Theory MSE RP BLUP: TMSE4	Theory MSE RP+re BLUP: TMSE5
20000	Estimates	8.3543408	3.3978577	3.4544560	3.3643140	3.3643140
20000	Theoretical	8.2910292	3.4033653	3.4607174	3.3709555	3.3709555
20000	Difference	0.0633116	-.0055076	-.0062614	-.0066415	-.0066415

Source:ced03p26.sas C:\projects\cluster\programs 5/8/03 EJS
Table 26-1. Summary Population Parameters (Same as above)

Table 26-2. Summary of MSE results: N=25 (n=20) M=10 (m=8) Pop-Mean=50 Resp.Error-Var=0

Total # of Simulated Clusters: NTOT	Statistic Parameter Differnce: TYP	Theory Cluster Mean: TMSE1	Theory MSE MM BLUP: TMSE2	Theory MSE SS BLUP: TMSE3	Theory MSE RP BLUP: TMSE4	Theory MSE RP+re BLUP: TMSE5
20000	Estimates	0.5072368	0.7903082	0.4656263	0.4626868	0.4626868
20000	Theoretical	0.5181893	0.7929697	0.4752844	0.4718099	0.4718099
20000	Difference	-.0109525	-.0026615	-.0096581	-.0091231	-.0091231

We plan to report results in the manuscript similar to Table 1. We have not yet constructed this table, and may revise the table's format based on our results. At this point, we have not evaluated the predictors based on estimated variances, nor have we corrected predictors that include response error.

Table 1. MSE based on Simulation of Predictors of Realized Cluster Means with n=300, m=3

			Response Error 0			Response Error 10v			Response Error 30v		
			M			M			M		
		Predictors	7	30	365	7	30	365	7	30	365
		P(MM)									
		P(SS)									
	Level 1	P(RP)									
		E-P(MM)									
		E-P(SS)									
		E-P(RP)									
		P(MM)									
		P(SS)									
Between Cluster Variance (v)	Level 2	P(RP)									
		E-P(MM)									
		E-P(SS)									
		E-P(RP)									
		P(MM)									
		P(SS)									
	Level 3	P(RP)									
		E-P(MM)									
		E-P(SS)									
		E-P(RP)									