Predicting Realized Random Effects with Clustered Samples from Finite Populations with Response Error
Edward J. Stanek III
Department of Biostatistics and Epidemiology, UMASS, Amherst, MA
Julio Singer
Department of Statistics, U. of São Paulo, Brazil

ABSTRACT

Finite clustered populations frequently occur where there is interest in parameters for the population and individual clusters. We develop methods for predicting linear combinations of population values such as realized cluster means based on a two-stage sampling probability model when response error is present. The predictors are closely related to best linear unbiased predictors (BLUP) that arise from common mixed model methods, as well as being closely related to model-based prediction approaches that postulate a super-population in survey sampling. The development assumes clusters are of equal size with equal size sampling of within cluster units. As in other survey prediction approaches, predictors are obtained for any linear combination, including clusters not realized in the sample.

While results are closely related to other methods, aspects of the development make this framework particularly appealing. The two stage-sampling model is directly used to develop predictors and standard errors, with no additional assumptions required. For a realized cluster, the interpretation of predictors can be directly related to predictors of realized second stage units (SSUs), and predictors of SSUs not realized, accounting directly for the second stage sampling fraction. With this interpretation, commonly used BLUP estimates can be seen to predict only the un-observed second stage units, not the realized cluster mean. The development reveals that two-stage sampling does not give rise to the variance structure assumed in super-population models, and thus provides interpretable alternatives to an apparently artificial model based approach. With response error present, we distinguish between target random variables that correspond to realized latent cluster values, and realized average cluster response and distinguish predictors between the two.
Contact Address:
Edward J. Stanek III
Department of Biostatistics and Epidemiology
404 Arnold House
University of Massachusetts at Amherst
Amherst, Ma. 01002

Phone: 413-545-3812
Fax: 413-545-1645
Email: stanek@schoolph.umass.edu

KEYWORDS: mixed models, super-populations, best linear unbiased predictors, cluster sampling, response error, finite populations, random permutation, optimal estimation, prediction, random effects, design based, model based, inference

ACKNOWLEDGEMENT. Many people contributed to the development of ideas that lead to this manuscript, including Dalton Andrade, Heleno Bolfarine, John Buonaccorsi, Viviana Lencina, and Oscar Louriero. This work was developed with the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), FINEP (PRONEX), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil and the National Institutes of Health (NIH-PHS-R01-HD36848), USA.
1. INTRODUCTION

Prediction of realized random effects is important in biological and environmental applications. The increased importance has resulted from popularization of mixed model methods and software (such as SAS PROC MIXED, Littell et al. (1996)) for analyzing clustered data, along with the assertion that the predictors of random effects may have useful interpretation. Best linear unbiased predictors (BLUP) are typically used for realized random effects based on mixed model theory (Goldberger 1962; Henderson 1984; McLean, Sanders et al. 1991; Robinson 1991; Stanek et al., 1999; McCulloch and Searle 2001). The sample data are typically considered to have been selected from a conceptual infinite population, or motivated as arising from an assumed Bayesian model.

Often, the underlying study populations are nested and finite, and measurement error is present. We present frequentist based methods for predicting linear combinations of fixed and realized random effects in two stage cluster sampling that account for finite population size and measurement error. The approach closely parallels the prediction based approach used with super-population models in survey sampling (Scott and Smith 1969; Bolfarine and Zacks 1992; Valliant, Dorfman et al. 2000). Instead of assuming a probability model connecting the population to a super-population, we directly represent the sample via a two stage random permutation model for the finite population. Target parameters correspond to the latent value of a potentially realized cluster.

In a Bayesian context, the true value of a parameter is considered to be the realized value of a random variable (Bickel and Doksum, 1977). Our goal is prediction of such a realized random variable, and in this sense, the results have a Bayesian interpretation. In the context of two stage cluster sampling, this definition of a potentially realized cluster is not unique. The target parameter underlying the random variable may be the latent value of a cluster (equal to the expected mean for the cluster). Alternatively, the target parameter may be the average response for the cluster. The two target parameters differ by the average response error.

These differences in target parameters can be distinguished in the finite population permutation model framework. We develop predictors for each. While there may be common interest in predicting the latent value of a potentially realized cluster, we show that the target parameters underlying the predictors of random cluster means from both mixed models and super-population models do not correspond to these latent values. Rather, the target parameter corresponds to the average response for the cluster, and includes average response error.

To help fix ideas, we first define the basic context and notation. Our development is limited to settings where each cluster is the same size, and equal size samples are selected from each selected cluster. Mixed model and prediction based super-population model results are briefly reviewed using the common notation. The main results are then developed for a two-stage random permutation model. We conclude with an illustration and some summary observations.

1.1 Basic Context and Notation

Let a finite population be defined by a listing of \( t = 1, \ldots, M \) subjects in each of \( s = 1, \ldots, N \) clusters, where the \( k^{th} \) measure of response for subject \( t \) in cluster \( s \) is given by

\[ Y_{stk} = Y_{st} + W_{stk} \quad (1.1) \]
The model for subject \( t \) in cluster \( s \) is a measurement error model where \( y_{st} \) is a fixed constant representing the expected response for the subject, and \( W_{st} \) represents measurement error (with zero expected value). We assume that a two stage cluster sample is to be selected from this population (using without replacement sampling), with \( m \) subjects selected for each of \( n \) clusters. On each selected subject, we assume that a single measure is to be made of response.

We represent the sample of clusters by defining a second list that is a random permutation of the clusters in the first list. A new label, \( i = 1, \ldots, N \), is assigned to the positions corresponding to different clusters the list. Without loss of generality, we assume that the first \( i = 1, \ldots, n \) positions contain the clusters that will occur in the sample. In a similar manner, we distinguish the listed subjects in cluster \( s \) from a random permutation of subjects in cluster \( s \), whose positions are labeled by \( j = 1, \ldots, M \). Without loss of generality, we assume that the first \( j = 1, \ldots, m \) positions contain the subjects that will occur in the sample if cluster \( s \) is selected. Using this notation we represent the population under a two stage permutation as an ordered list of \( NM \) random variables, where a model for the subject in the \( j^{th} \) position in the cluster in the \( i^{th} \) position is represented by

\[
y^*_{ijk} = y_j + W^*_{jk}. \tag{1.2}
\]

For ease of exposition, we refer to the cluster that will occupy position \( i \) in the permutation as the \( i^{th} \) primary sampling unit (PSU), and to the subject that will occupy position \( j \) in the permutation as the \( j^{th} \) secondary sampling unit (SSU). Thus, \( Y_j \) represents the expected response for the \( j^{th} \) SSU in the \( i^{th} \) PSU.

Prior to sample selection, the cluster that will correspond to PSU \( i \) is random, and hence the expected value of response for the cluster in the \( i^{th} \) position is a random variable. Once the sample has been selected, it will be apparent which cluster corresponds to a particular position (PSU) in the sample. We refer to the cluster that results as the realized PSU, and the expected value of that cluster as the realized random effect (noting that such effects are often defined as deviations relative to an overall mean). In this context, different strategies may be advocated for predicting the expected value of a realized PSU. The objective of this paper is presentation of a new approach to this problem based directly on the probability models underlying finite population sampling. We first review alternative strategies for predicting a realized PSU mean using mixed models and super-population models in a balanced setting without response error, setting the stage for development of the main results in the subsequent section.

1.2 Mixed Models

In the mixed model literature, the model for the response of the \( j^{th} \) SSU in the \( i^{th} \) PSU is given by

\[
Y_j = \mu + B_i + E_j \tag{1.3}
\]

where \( \mu \) corresponds to the overall population mean, \( B_i \) is a random effect that corresponds to the deviation from the population mean for the \( i^{th} \) selected PSU, and \( E_j \)
is the random error corresponding to the deviation of the $j^{th}$ selected SSU from the $i^{th}$ selected PSU mean. Typically, additional assumptions are made. The random effects, $B_i$, are assumed independent and identically normally distributed, $B_i \sim iid \ N(0, \sigma^2)$, and independent of the random errors, also assumed independent and identically distributed for all $j$ in $i$, $E_{ij} \sim iid \ N(0, \sigma^2_{ij})$. The model and assumptions enable the joint distribution of fixed and random effects to be specified. This density, when maximized jointly with respect to the fixed and random effects, leads to Henderson’s mixed model equations (Henderson, Kempthorne et al. 1959).

In model (1.3), assuming known variances, the estimator of the fixed effects is the weighted least squares estimator, $$\hat{\mu} = \sum_{i=1}^{n} w_i \bar{Y}_i$$ where $$\bar{Y}_i = \frac{\sum_{j=1}^{m} Y_{ij}}{m}, \quad w_i = \frac{1/v_i}{\sum_{i'=1}^{n} 1/v_{i'}}$$ and $v_i = \sigma^2 + \frac{\sigma^2_i}{m}$. Solution of the mixed model equations results in the BLUP of $B_i$.

Combining the estimator of $\mu$ and the predictor of $B_i$, the predictor of $\mu + B_i$ is given by

$$\hat{p} = \hat{\mu} + k_i (\bar{Y}_i - \hat{\mu}) \quad (1.4),$$

where $k_i = \frac{\sigma^2}{v_i}$.

In practice, the variance parameters are replaced by the maximum likelihood or restricted maximum likelihood estimates. McCulloch and Searle (2001) note that there are many other derivations of the BLUP estimates (discussed extensively by Robinson (1991)). A derivation of particular note is the Bayesian derivation, where (1.4) is the expected value of the posterior distribution, assuming normality. The normality assumptions are not necessary, and can be replaced by knowledge of the first and second moments. None of these derivations or discussions account for the impact of a finite population on the predictors.

1.3 Super-population Sampling Models

Predictors of a linear combination of elements of a finite population in a two stage sampling setting was developed by Scott and Smith (1969) using a super-population model. In their approach, a finite population is viewed as the realization of a set of random variables. We refer to the set of random variables as the super-population, and specify a model for the super-population. The parameter of interest is a linear combination of values in the finite population, such as the average response for a PSU. Such parameters can be defined for all PSUs in the finite population. A sample will result in the realization of part of the finite population. The essential statistical problem is how best to estimate the linear combination of values of the remaining random variables that are not realized. Estimates are constructed using the joint distribution assumed for the super-population. The result is called a predictor, since it is composed in part of predicted values of un-observed random variables (Royall 1976; Bolfarine and Zacks 1992; Valliant, Dorfman et al. 2000). This approach to statistical inference in
survey sampling is called model based, since inference is based on the model assumed for the super-population.

We define a nested super-population model by the first and second moments. The random variables \( Y_i, i = 1, ..., N; j = 1, ..., M \) in the super-population are assumed satisfy

\[
E(Y_{ij}) = \mu \quad \text{and} \quad \text{cov}(Y_{ij}, Y_{il}) = \delta^2 + \sigma_i^2 \quad \text{when} \quad i = k, j = l
\]

\[
= \delta^2 \quad \text{when} \quad i = k, j \neq l
\]

\[
= 0 \quad \text{otherwise.}
\]

Scott and Smith (1969) used this super-population model to derive predictors of linear combinations of elements in the population that correspond to a PSU mean based on two stage cluster sampling. Scott and Smith frame the statistical problem as prediction of the remaining un-observed SSUs in the realized PSU. A Bayesian derivation (assuming the super-population is normally distributed), and a distribution free derivation (based on minimizing the expected mean squared error (MSE) of a linear predictor) is given, and shown to result in the same predictor given by

\[
\hat{P} = \frac{m}{M} \bar{Y}_t + \left( \frac{M-m}{M} \right) \left[ \hat{\mu}^* + k_i^* (\bar{Y}_t - \hat{\mu}^*) \right] \quad (1.5)
\]

if the PSU is partially realized in the sample, or \( \hat{P} = \hat{\mu}^* \), if the PSU is not included in the sample, where \( \hat{\mu}^* = \sum_{i=1}^{M} w_i^* \bar{Y}_i \) where \( w_i^* = \frac{1}{\sum_{j=1}^{M} \sigma_i^2}, \quad v_i^* = \delta^2 + \frac{\sigma_i^2}{m}, \quad \text{and} \quad k_i^* = \frac{\delta^2}{v_i^*} \). The predictor (1.5) has a nice interpretation as the weighted sum of two terms: the sample mean for a realized PSU, and predictors of the remaining SSUs for the realized PSU. The weighting factor is the proportion of SSUs that are observed and the proportion that are to be predicted. If none of the SSUs are realized, the predictor simplifies to the weighted sample mean.

There is an obvious similarity between equation (1.4) and equation (1.5). When \( \frac{m}{M} \) is small enough that the first term in equation (1.5) can be ignored, and \( 1 - \frac{m}{M} \approx 1 \), the two equations appear to be identical, with the exception of different representations for the variance components. In this setting, one can interpret the best linear unbiased predictor (in equation (1.4)) as a predictor of the unobserved SSUs in a realized PSU. If the number of SSUs in the realized PSU is so large that the observed SSUs are a negligible fraction of the total for the PSU, then the realized PSU mean is estimated by the predicted values of the unobserved SSUs for the PSU. When a non-trivial portion of the SSUs are observed in the sample for a realized PSU, then the estimate of the realized PSU mean is a weighted average of the mean of the realized SSUs, and predicted values of the unobserved SSUs for the PSU. This provides a strong intuitive appeal to the prediction based approach as advocated by Valliant, Dorfman, and Royall (2000).

As discussed by Scott and Smith, the assumption of normality is not needed, and can be replaced by the assumption that PSUs are exchangeable, and SSUs are exchangeable within PSUs. The predictor is assumed to be a linear function of the sample responses. Coefficients of the predictor are derived so to minimizes the expected
MSE over the super-population. The MSE is required to be bounded, leading to a constraint in the minimization.

Scott and Smith’s result appears to be of direct relevance to two stage cluster sampling. The context is not the same as repeated two stage sampling of a fixed finite population. The difference is subtle. Results hold for a setting where a finite population is selected from the super-population, and then two stage samples are selected from the realized population, with the whole process repeated. The addition of the super-population introduces another step in the sampling. Since the super-population is defined only in terms of moments of random variables, some artificiality is introduced. This artificiality has been argued by some to be a strength (Cochran (1946)) when the super-population represents the collection of repeated snap-shots of the finite population, perhaps at different times. We develop the predictor of a PSU mean directly from a model based on repeated two stage sampling from a fixed finite population. The development does not require postulating a super-population.

2. THE TWO STAGE PERMUTATION MODEL FOR A FINITE POPULATION

We derive predictors of a PSU mean from a two stage clustered sample of a finite population. The probability model corresponds to the two stage permutation of clusters and subjects in the population. The development closely parallels that used in the prediction based approach to survey sampling (Royall 1976). A principal distinction is the use of a probability model based directly on the two stage sampling, and retaining the identifiability of the population units. The results, while similar in form to those of Scott and Smith (1969) and Vallient, Dorfman, and Royall (2000), do not require positing a super-population. Assumptions are simple and transparent.

In this section, we define the population and target parameters, and notation. Next we define the random variables representing a two stage random permutation of the population, and re-arrange these random variables into a sampled and remaining set. We assume equal size clusters and equal size sampling within clusters, and allow for response error.

2.1. Finite Population Parameters and Re-parameterizations

We begin by defining parameters for the population represented by model (1.1). First, we note that the expected value of $W_{stk}$ is zero for all $s=1,...,N$ and $t=1,...,M$, and define the response error variance for subject $t$ in cluster $s$ as $\sigma^2_{st}$. The average response error variance we denote by

$$\bar{\sigma}^2 = \sum_{s=1}^{N} \sum_{t=1}^{M} \frac{\sigma^2_{st}}{NM}.$$ (2.1)

The expected value of response for subject $t$ in cluster $s$ is represented by the fixed constant $y_{st}$. We define the mean and variance of the expected response for subjects in cluster $s$ as

$$\mu_s = \frac{1}{M} \sum_{t=1}^{M} y_{st} \text{ and } \left(\frac{M-1}{M}\right)\sigma^2_s = \frac{1}{M} \sum_{t=1}^{M} (y_{st} - \mu_s)^2$$
for \( s = 1, \ldots, N \) (where we use the survey sampling definition of the parameter \( \sigma_s^2 \)). We define the average within cluster variance as

\[
\sigma_s^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2. \tag{2.2}
\]

Similarly, the overall mean, and the variance between cluster means is defined as

\[
\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s \quad \text{and} \quad \sigma^2 = \frac{1}{N} \sum_{s=1}^{N} (\mu_s - \mu)^2.
\]

Using these parameters, we represent the potentially observable response for subject \( t \) in cluster \( s \) as

\[
y_{st} = \mu + \beta_s + \varepsilon_{st} \tag{2.3}
\]

where \( \beta_s = (\mu_s - \mu) \) is the deviation of the mean for cluster \( s \) from the overall mean, and \( \varepsilon_{st} = (y_{st} - \mu_s) \) is the deviation of subject \( t \)'s response (in cluster \( s \)) from the mean for cluster \( s \). Model (2.3) is called a derived model (Hinkelmann and Kempthorne 1994).

Defining \( y = \left( y_1', y_2', \ldots, y_N' \right)' \) where \( y_s = (y_{s1}, y_{s2}, \ldots, y_{sM})' \), model (2.3) can be summarized as

\[
y = X\mu + Z\beta + \varepsilon \tag{2.4}
\]

where

\[
X = I_N \otimes 1_M, \quad \tag{2.5}
Z = I_N \otimes 1_M, \quad \tag{2.6}
\beta' = (\beta_1, \beta_2, \ldots, \beta_N), \quad \tag{2.7}
\]

\( 1_a \) is an \( a \times 1 \) column vector of ones, \( \otimes \) denotes the Kronecker product (Graybill 1983), and \( \varepsilon \) is defined similarly to \( y \). None of the terms in model (2.4) are random variables. Defining \( W = \left( W_{st} \right) \) similar to \( y \), and adding it to \( y \) summarizes the stochastic model given by (1.1). Such a model is a response error model for all elements in the finite population.

2.2. Random Variables and The Two Stage Random Permutation Model

We define a stochastic model that parallels model (2.4) for two stage cluster sampling. A sample of clusters can be identified by the first \( n \) clusters in a random permutation where we label the positions in the permutation as \( i = 1, \ldots, N \) with the sample PSUs corresponding to \( i = 1, \ldots, n \). The basic units permuted are the vectors, \( y_s, s = 1, \ldots, N \). Similarly, a sample of subjects in a cluster can be identified by selecting the first \( m \) subjects in a permutation of subjects in a cluster. We use this idea to define random variables whose realization corresponds to a two stage permutation of the population. Assuming each realization of the permutation is equally likely, the random variables formally represent two stage sampling (Cochran 1977).

We use different terms and indices for the finite population and the permuted population. In the finite population, we index clusters by \( s = 1, \ldots, N \) and subjects by \( t = 1, \ldots, M \). In a two stage permutation of the finite population, we refer to the
permuted clusters as primary sampling units (PSUs), and the permuted subjects in a PSU as the secondary sampling units (SSUs). We index the position of the PSUs in the permutation by \( i = 1, \ldots, N \) and the position of the SSUs by \( j = 1, \ldots, M \). In the finite population, we represent the expected response for subject \( t \) in cluster \( s \) by \( y_{st} \); in the permuted population, we represent the random variable for response of the \( j^{th} \) selected SSU in the \( i^{th} \) selected PSU by \( Y_{ij} \).

We define sampling indicator random variables to relate \( y_{st} \) to \( Y_{ij} \). The indicator random variable \( U_{is} \) takes on a value of one when the \( i^{th} \) PSU is cluster \( s \), and a value of zero otherwise; the indicator random variable \( U_{ij}^{(s)} \) takes on a value of one when the \( j^{th} \) SSU in cluster \( s \) is subject \( t \), and zero otherwise. Using these random variables, the random variable corresponding to the \( i^{th} \) PSU and the \( j^{th} \) SSU in a permutation is given by

\[
Y_{ij} = \sum_{s=1}^{N} \sum_{t=1}^{M} U_{is} U_{ij}^{(s)} y_{st} \quad (2.8).
\]

Defining \( U^{(s)} = (U_{1s}^{(s)} \ U_{2s}^{(s)} \ \cdots \ U_{Ms}^{(s)}) \) where \( U_{is}^{(s)} = (U_{1s}^{(s)} \ U_{2s}^{(s)} \ \cdots \ U_{Ms}^{(s)})' \) is an \( M \times 1 \) column vector, and \( U = (U_1 \ U_2 \ \cdots \ U_N)' \), where \( U_s = (U_{1s} \ U_{2s} \ \cdots \ U_{Ns})' \) is an \( N \times 1 \) column vector, the random variables are given by

\[
Y = (U \otimes I_M) \left( \bigoplus_{s=1}^{N} U^{(s)} \right) y \quad (2.9)
\]

where \( Y = \left( (Y_{ij}) \right) = (Y_1' \ Y_2' \ \cdots \ Y_N')' \) and \( Y_i = (Y_{i1} \ Y_{i2} \ \cdots \ Y_{im})' \), and \( \bigoplus A_s \) denotes a block diagonal matrix, with blocks given by \( A_s \) (Harville 1997). In a similar manner, we permute the vector response errors, defining

\[
W^* = (U \otimes I_M) \left( \bigoplus_{s=1}^{N} U^{(s)} \right) W. \quad (2.10)
\]

Adding (2.9) to (2.10) results in

\[
Y^* = Y + W^*. \quad (2.11)
\]

This is the two stage random permutation model with response error that summarizes the stochastic model for random variables representing the finite population given by (1.2).

We use the re-parameterized model (2.4) and relationship (2.9) to express the random permutation model as a mixed model. Using the elementary properties of the indicator random variables and (2.5) and (2.6), \( U_1 = 1_N \ ; \ U_{1}^{(s)} I_M = 1_M \).
\((U \otimes I_M)^{(N)} X = X\) and \((U \otimes I_M) \left( \oplus_{i=1}^{N} U^{(i)} \right) Z = U \otimes I_M = ZU\). We define

\[
E = (U \otimes I_M) \left( \oplus_{i=1}^{N} U^{(i)} \right) \varepsilon \quad \text{and}
\]

\[
B = U\beta 
\]

where \(B = (B_1 \ B_2 \ \cdots \ B_N)\). The terms \(B_i = \sum_{s=1}^{N} U_{is} \beta_s\) for \(i = 1, \ldots, N\) in (2.12) are random effects, and represent the deviation from the overall mean for the mean of the cluster in the \(i^{th}\) position in the permutation. Combining these expressions, we obtain the random permutation model

\[
Y^* = X\mu + ZB + (E + W^*) \quad . \tag{2.13}
\]

Model (2.13) differs from the usual mixed model since it represents all elementary units in the population, as opposed to solely sampled units. The last terms in the model, \((E + W^*)\), represent the permutation of deviations of SSUs about the PSU mean, and the corresponding measurement error. The average of the first \(M\) random variables in \(Y\) is the mean for the first cluster in the permutation. A realization of \(Y\) represents a particular two-stage permutation of the population.

We assume that there is interest in a linear combination of random variables,

\[
T = g'Y \quad \tag{2.14}
\]

such that \(g' = (g_1' \ g_2' \ \cdots \ g_N')\), where \(g_i = \left( \begin{array}{c} g_{ii} \\ \end{array} \right)\) is an \(M \times 1\) vector of constants. Linear combinations of PSU means may be defined by

\[
g' = \frac{1}{M} b' \otimes 1_M' \quad \tag{2.15}
\]

where \(b = \left( \begin{array}{c} b_i \end{array} \right)\) is an \(N\)-dimensional column vector of constants. Of principal interest is the linear combination that defines the mean of a PSU in position \(i\), such that

\[
g' = \frac{1}{M} e_i' \otimes 1_M' \quad . \tag{2.16}
\]
with $e_i$ defined as $N$-dimensional column vector with a value of one in the $i^{th}$ position, and zero elsewhere. Using (2.16),

$$T = \sum_{s=1}^{N} U_{is} \mu_s . \ (2.17)$$

In the context of the super-population model described by Scott and Smith (1969) that was extended to include response error by Bolfarine and Zacks (1992), linear combinations of the random variables are defined such that the mean of a PSU in position $i$ is defined as

$$T_a = g^T \mathbf{Y}^* \ (2.18)$$

where $T_a = T + g^T \mathbf{W}^*$. The random variables represented by (2.14) and (2.18) have different interpretations, with the difference being the addition of a response error term added to the random variable defined using the super-population model framework. The ability to distinguish such subtleties in target random variables is a strength of the finite population sampling based approach.

Under the two stage random permutation model (2.13), using the subscript $\xi_1$ to denote expectation with respect to permutations of the clusters, the subscript $\xi_2$ to denote expectation with respect to permutations of subjects in a cluster, and the subscript $\xi_3$ to denote expectation with respect to response error,

$$E_{\xi_1 \xi_2 \xi_3} (\mathbf{Y}^*) = \mathbf{X} \mu \ \ (2.19)$$

and

$$\text{var}_{\xi_1 \xi_2 \xi_3} (\mathbf{Y}^*) = \frac{\bar{\sigma}^2}{N} (\mathbf{I}_N \otimes \mathbf{I}_M) + \sigma^2_e (\mathbf{I}_N \otimes \mathbf{P}_M) + \sigma^2 (\mathbf{P}_N \otimes \mathbf{J}_M)$$

$$= \left( \frac{\bar{\sigma}^2}{N} + \sigma^2_e \right) \mathbf{I}_{NM} + \sigma^2 (\mathbf{I}_N \otimes \mathbf{J}_M) - \frac{\sigma^2}{M} \mathbf{J}_{NM} \ \ (2.20)$$

where

$$\sigma'^2 = \sigma^2 - \frac{\sigma^2_e}{M} \ \ (2.21)$$

and $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a} \mathbf{J}_a$. (see Appendix A).

3. PREDICTING A PSU BASED ON A TWO STAGE SAMPLE

We assume that sampling results in a single measure on each SSU selected via a two stage simple random sample without replacement, with $m$ SSUs selected from each of $n$ selected PSUs. The process of sampling gives rise directly to the stochastic representation of the population in model (2.13). The random variables corresponding to the sample will be realized. Our principal interest is focused on predicting the expected value of a PSU defined by (2.14) and (2.16), but we develop the predictor for more general linear combinations of random variables.

We develop a predictor that is a linear function of the sample, is unbiased, and that minimizes the expected value of the mean squared error (MSE). The basic strategy is given in many places (Scott and Smith (1969); Royall (1976); Bolfarine and Zacks
(1992); Valliant, Dorfman et al. (2000)). We first partition the elements of $Y^*$ into a sampled and remaining portion 

$$\frac{Y'_i}{Y''_i} = \left( \frac{K'_i}{K''_i} \right)^{'} \text{ where } K_i = \left( I_n \mid 0 \right) \otimes \left( I_m \mid 0 \right).$$

and $K''_i = \left( I_n \mid 0 \right) \otimes \left( I_{M-n} \mid 0 \right) \otimes I_M$. We partition $X$ in a similar manner, such that

$$\frac{X'_i}{X''_i} = \left( \frac{K'_i}{K''_i} \right)^{'} \text{ and represent var}_{i,j,k} \left( \frac{Y'_i}{Y''_i} \right) = \text{var}_{i,j,k} \left( \frac{Y'_i}{Y''_i} \right) + \tilde{\sigma}^2 \left( I_{nm} \otimes 0 \right).$$

where $\text{var}_{i,j,k} \left( \frac{Y'_i}{Y''_i} \right) = \left( V_I \mid V_{I,II} \right)$. We also partition $g'$, resulting in

$$\left( g'_i \mid g''_i \right) = \left( g'K'_i \mid g'K''_i \right) \text{ and } g''_i = \left( g'_{I,II} \mid g'_{2,II} \right).$$

We assume that the subset $Y'_i$ of the elements of $Y^*$ will be realized, call the subset "the sample", and express $T$ (or $T_i$) as the sum of two parts, one depending on the potentially realized sample, and the other depending on the remaining random variables. We require the predictor of (2.14) to be a linear function of the sampled random variables such that

$$\hat{T} = \left( g' + a' \right) Y'_i \text{, to be unbiased such that } a'X_i - g'_{I,II}X_{I,II} = 0 \text{, and to minimize the expected value of the mean squared error given by } \text{var}_{i,j,k} \left( \hat{T} - T \right).$$

The best linear unbiased predictor is developed using Theorem 2.1 (Royall 1976). Using the expression

$$\hat{T} - T = \left( \left( a' + g'_i \right) \mid -g'_i \mid -g''_i \right) \underbrace{\left( \begin{array} {c} Y'_i \\ E_{i,j,k} \left( Y'_i \right) \\ E_{i,j,k} \left( Y''_i \right) \end{array} \right)}_{\text{and noting that}}$$

$$\text{var}_{i,j,k} \left( \left( \begin{array} {c} Y'_i \\ E_{i,j,k} \left( Y'_i \right) \\ E_{i,j,k} \left( Y''_i \right) \end{array} \right) \right) = \left( \begin{array} {ccc} V_I & V_I & V_{I,II} \\ V_I & V_I & V_{I,II} \\ V_{II,II} & V_{II,II} & V_{II} \end{array} \right)$$

where $V_I = V_I + \tilde{\sigma}^2 I_{nm}$, the predictor is given by

$$\hat{T} = g'_i \left[ X_i \hat{\alpha} + V_I V_I^{-1} \left( Y'_i - X_i \hat{\alpha} \right) \right] + g''_i \left( X_i \hat{\alpha} + V_{I,II} V_{I,II}^{-1} \left( Y'_i - X_i \hat{\alpha} \right) \right) \tag{3.1}$$
where \( \hat{\alpha} = \left( X'_i V^{-1}_r X_i \right)^{-1} X'_i V^{-1}_r Y_i \). Terms in this expression are given by

\[
\hat{\alpha} = \bar{Y} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} Y^*_{ij}, \quad V_i = \sigma^2 I_{nn} + \sigma^2 (I_n \otimes J_m) - \frac{\sigma^2}{N} J_{nn},
\]

\[
V_i^{-1} = \frac{1}{\left( \sigma^2 + \bar{\sigma}^2 \right)} \left( I_n \otimes \left( I_m - k^* \frac{J_m}{m} \right) \right) + \frac{k^* \sigma^2}{m \sigma^2 N \left( \sigma^2 + \bar{\sigma}^2 \right) + m \sigma^2 - \frac{nn \sigma^2}{N}} J_{nn}
\]

where

\[
k^* = \frac{m \sigma^2}{m \sigma^2 + \sigma^2 + \bar{\sigma}^2} \quad \text{and} \quad V_{i,ii} = \sigma^2 \left( I_n \otimes \frac{J_m}{(M-m)n_m} \right) - \frac{\sigma^2}{N} J_{nn} \quad \text{(3.2)}
\]

Details of this derivation are presented by Stanek (2002a). Using expressions for the partitioned variances, and defining \( \bar{Y}_i \), \( \bar{V}_i \), \( k_e \), \( k_m \), and \( k_e^* \), and the predictor simplifies to

\[
\hat{r} = g'_i \left[ k_e \left( I_n \otimes P_m \right) + \left( \left[ \frac{J_n}{n} + k_r P_n \right] \otimes \frac{J_m}{m} \right) \right] Y_i^* + g'_{1,ii} \left( \left[ \frac{J_n}{n} + k^* P_n \right] \otimes \frac{(M-m)n_m}{m} \right) Y_i^* + g'_{2,ii} \left( \frac{J_n}{n} \otimes \frac{J_{nm}}{m} \right) Y_i^* \quad \text{(3.2)}
\]

The first term in equation (3.2) is the predictor of the linear combination of the realized SSUs for the PSUs that occur in the sample. The second term is the predictor of the linear combination of the remaining SSU’s for the sample PSUs. The last term is the predictor of the linear combination of the SSUs for PSUs that are not in the sample. Defining \( g' \) as in (2.16), the predictor in (3.2) simplifies to

\[
\hat{r} = \left[ \frac{m}{M} \right] e'_{ir} \left( I_n \bar{V}^* + k_r P_n \bar{V}_i^* \right) + \left[ \frac{M - m}{M} \right] e'_{ir} \left( I_n \bar{V}^* + k^* P_n \bar{V}_i^* \right), \quad \text{(3.3)}
\]

where \( e'_{ir} = \left( e'_{ir} \right. \left. \begin{array}{c} e'_{ir} \\ e'_{ir} \end{array} \right) \). The term on the first line of (3.3) is the predictor of a PSU that is realized in the sample (when there is interest in a PSU in position \( i \), where \( i \leq n \)); the term on the second line is the predictor of a PSU that is not realized in the sample (when there is interest in a PSU in position \( i > n \) ). The term \( k^* \) accounts for response error,
and shrinks the realized values of SSUs in a sample towards the sample mean when a sample PSU is of interest. The term \( k^* \) accounts for both response error and between SSU variations, and shrinks the predictors of the SSUs not included in the sample towards the sample mean. A similar result holds for more general linear combinations of PSU means where \( e'_{i} = \left( \begin{array}{c} e'_{i,1} \\ e'_{i,2} \\ \vdots \\ e'_{i,n} \end{array} \right) \) is replaced by \( b'_{i} = \left( \begin{array}{c} b'_{i,1} \\ b'_{i,2} \\ \vdots \\ b'_{i,N-n} \end{array} \right) \) in (3.3).

The predictor of (2.14) given by (3.2) can be compared to the predictor of (2.18). A similar derivation results in the best linear unbiased predictor of (2.18) given by

\[
\hat{T}_a = g'_1 Y^*_i + \left[ g'_{1,if} \left( \frac{J_n + k^* P_n}{n} \right) \otimes \frac{J}{m} \right] Y^*_i + \left[ g'_{2,if} \left( \frac{J_{(N-n)n}}{n} \otimes \frac{J}{m} \right) \right] Y^*_i
\]

which simplifies (when \( g' \) is defined by (2.16)) to

\[
\hat{T}_a = \left( \frac{m}{M} \right) e'_{i} Y^*_i + \left( \frac{M - m}{M} \right) e'_{i} \left( 1_n \bar{Y}^* + k^* P_n \bar{Y}_i^* \right) + \left[ e'_{i,1} 1_{N-n} \right] \bar{Y}^*.
\]

This predictor is nearly identical to the predictor given by Bolfarine and Zacks (1992), apart from a similar difference in PSU variance parameter definitions. Since the target random variable defined by (2.14) includes response error, the values of the realized SSUs are used directly in the predictor.

When there is no response error, the target random variable given by (2.14) and (2.18) are the same. Since \( \sigma^2 = 0 \), and the predictor given by (3.3) (or (3.5)) simplifies to

\[
\hat{T} = \left( \frac{m}{M} \right) e'_{i} Y^*_i + \left( \frac{M - m}{M} \right) e'_{i} \left( 1_n \bar{Y} + k P_n \bar{Y}_i \right) + \left[ e'_{i,1} 1_{N-n} \right] \bar{Y}
\]

since \( e'_{i} \left( 1_n \bar{Y} + P_n \bar{Y}_i \right) = e'_{i} \bar{Y}_i \), where \( \bar{Y}_i = \left( \sum_{j=1}^{m} Y_{ij} \right) / m \), \( \bar{Y}_j = \left( \bar{Y}_1 \bar{Y}_2 \cdots \bar{Y}_n \right)' \), \( \bar{Y} = \sum_{j=1}^{m} \sum_{j=1}^{m} Y_{ij} / nm \), and \( k = \frac{m \sigma^2}{m \sigma^2 + \sigma_e^2} \). This result is nearly identical to the predictor developed by Scott and Smith (1969), with the exception that the shrinkage constant is defined relative to a between PSU variance component given by (2.21) as opposed to \( \delta^2 \). Although the predictors given by (1.5) and (3.6) are nearly identical, the context is very different, since no super-population is required for (3.6).

3.1. The MSE of the Predictor

We develop expressions for the MSE of the best linear unbiased predictors, using the target random variables defined by (2.14) or (2.18), with predictors given by (3.2) or (3.4). For ease of exposition, we express (3.2) as
\[ \hat{T} = \left[ g'_1 C'_1 + g'_{1,II} C'_2 + g'_{II} C'_3 \right] Y_I^* \]  

(3.7)

where \( C'_1 = k_e \left( I_n \otimes P_m \right) + k_r \left( P_n \otimes \frac{J_{mm}}{m} \right) + \frac{J_{nm}}{nm} \); \( C'_2 = k_e \left( P_n \otimes \frac{J_{m,nm}}{m} \right) \); \( C'_3 \equiv \frac{J}{nm} \).

In a similar manner, \( \hat{T}_A = \left[ g'_1 + g'_{1,II} C'_2 + g'_{II} C'_3 \right] Y_I^* \). Using this notation, the expected MSE is given by

\[ \operatorname{var}_{\xi_{1,2,3}} (\hat{T}) = C + D \]  

(3.8)

or

\[ \operatorname{var}_{\xi_{1,2,3}} (\hat{T}_A) = C + \sigma^2 \sigma g'_{II} \]  

(3.9)

where

\[ C = g'_{1,II} \left[ C'_2 \left( V_I + \bar{\sigma} I_{nm} \right) C_2 \right] g_{1,II} + 2 g'_{II} \left[ \left( C'_3 \left( V_I + \bar{\sigma} I_{nm} \right) - V_{II,I} \right) C_2 \right] g_{1,II} \]

\[ + g'_{II} \left[ C'_3 \left( V_I + \bar{\sigma} I_{nm} \right) C_3 - 2 V_{II,I} C_3 + V_{II} \right] g_{II} \]

(3.10)

and

\[ D = g'_{1,I} \left[ (C'_1 - I_{nm}) V_I (C'_1 - I_{nm}) + \sigma^2 C'_1 C'_1 \right] g_{1,I} \]

\[ + 2 g'_{1,II} \left[ C'_2 V_I (C'_1 - I_{nm}) + \sigma^2 C'_2 C'_1 \right] g_{1,II} \]

\[ + 2 g'_{II} \left[ (C'_3 V_I - V_{II,I}) (C'_1 - I_{nm}) + \sigma^2 C'_3 C'_1 \right] g_{II} \]

(3.11)

For the special case where \( g' \) is defined by (2.15), (see Stanek, 2002b, p22 equation 1.11)

\[ C = \left[ \sigma^2 \left( 1 - k^* \right) + \frac{\sigma_e^2}{M} \right] \left( M - m \right) b'_t P_n b_{II} \]

\[ + \frac{1}{nn} \left[ \left( M \right) \left( \sigma_e^2 + \bar{\sigma}^2 \right) \left( M - m \right) b'_t J_n b_I \right] \]

\[ + \sigma^2 e'_{III} P_{N-n} e_{III} + \frac{1}{n} \left[ \sigma^2 + \frac{m \sigma^2}{M} + \sigma_2^2 \right] b'_{II} J_{N-n} b_{II} \]

\[ + \left[ \frac{2 \sigma^2}{nm} \right] \left( M - m \right) b'_t J_{N-N-n} b_I \]

and

\[ D = \frac{\bar{\sigma}^2}{M} \left[ \left[ \left( k^* (1 - k^* \right) \left( 1 - k^* \right) - k^* \right) \left( M - m \right) \right] b'_t P_n b_I \]

\[ + 2 k^* \left[ k^* - k^* - k_e \left( 1 - k^* \right) \right] \left( M - m \right) \]

\[ + \frac{1}{n} \left[ 2 \left( M \right) b'_t J_n b_I + \frac{2}{n} b'_{II} J_{N-n-n} b_I \right] \]

and...
\( \bar{\sigma^2} g'_i g''_i = \frac{\bar{\sigma^2}}{M} \left[ \left( \frac{M-m}{M} \right) b'_i b''_j + b'_i b''_j \right] \).

When \( g' \) is defined by (2.16), these expressions simplify further (Stanek, 2002b, p24 equation 1.13) when \( i \leq n \)

\[
\text{var}_{\tilde{\xi}_{i} \tilde{\xi}_{i}} \left( \hat{T} \right) = \\
\left( \frac{M-m}{M} \right)^2 \left[ \sigma^2 (1-k^*) + \frac{\sigma^2_e}{M-m} \right] \left( 1-\frac{1}{n} \right) + \frac{1}{nm} \left[ \left( \frac{M}{M-m} \right) \sigma^2_e + \bar{\sigma^2} \right] + \\
+ \frac{\bar{\sigma^2}}{M} \left[ k^* (1-k^*) + (1-k^*) (1-k^*) k_e + k_e^* \right] \left( \frac{m}{M} \right) \left( 1-\frac{1}{n} \right) + \frac{1}{n} \left[ 2 - \left( \frac{m}{M} \right) \right]
\]

and when \( i > n \),

\[
\text{var}_{\tilde{\xi}_{i} \tilde{\xi}_{i}} \left( \hat{T} \right) = \sigma^2 \left( 1-\frac{1}{N-n} \right) + \frac{\sigma^2_e}{n} + \frac{m \sigma^2 + \sigma^2_e + \bar{\sigma^2}}{nm},
\]

while the MSE for \( T_A \) can be developed in a similar manner, resulting in when \( i \leq n \)

\[
\text{var}_{\tilde{\xi}_{i} \tilde{\xi}_{i}} \left( \hat{T}_A \right) = \left( \frac{M-m}{M} \right)^2 \left[ \sigma^2 (1-k^*) + \frac{\sigma^2_e}{M-m} \right] \left( 1-\frac{1}{n} \right) + \\
+ \frac{\bar{\sigma^2}}{M} \left( \frac{M-m}{M} \right)
\]

and when \( i > n \),

\[
\text{var}_{\tilde{\xi}_{i} \tilde{\xi}_{i}} \left( \hat{T}_A \right) = \sigma^2 \left( 1-\frac{1}{N-n} \right) + \frac{\sigma^2}{n} + \frac{m \sigma^2 + \sigma^2_e + \bar{\sigma^2}}{nm} + \frac{\bar{\sigma^2}}{M}.
\]

In practice, variance parameters will be unknown and typically need to be estimated from the sample data. One practical strategy for estimation is use of restricted maximum likelihood, assuming normality, making use of the algorithms in computer programs such as Proc Mixed. In the balanced setting, the resulting estimates are method of moments, and hence do not require the normality assumption. Replacing variance parameters by estimates in the expressions the predictors is similar to an Empirical Bayes approach. The MSE will underestimate the true MSE since it will not account for variability in the variance estimates both in the predictor and in the MSE.

### 4. EXAMPLE

We consider a simple example from the a study of seasonal variation in serum cholesterol which we refer to as the Seasons study (Seasonal Variability of Blood Lipids,
NHLBI, number R01-HL52745, Merriam et al., 1999) to illustrate the methods. Four or more fasting serum cholesterol measures (with at least one in the summer and winter) were collected on a volunteer sample of \( n = 476 \), 20-70 year old members of the Fallon Health Maintenance Organization (HMO) supplemented by special minority based on 5000 patient contacts (Ockene et al, 2002). Participants were enrolled between 1994-1998. For each participant, measures were separated by consecutive three month intervals over a 12 month period. Assuming the 20-70 year old Fallon HMO membership is 25,000, we estimate the volunteer population to be of size \( N = 2380 \). Neither the study subjects (which we consider here as clusters), nor the time of measurement was random.

We construct predictors of average cholesterol for potentially realized subjects assuming a two stage random permutation model. We use the first four serum cholesterol measures for subjects with more than four measures. The model assumes the study subjects with four (or more) measures of serum cholesterol are a simple random sample from the potential volunteer subjects at the Fallon HMO, and that fasting serum cholesterol measures are made on a simple random sample of \( m = 4 \) days during a one year (\( M = 365 \) day) period. Only one measure of serum cholesterol was made on each sample day, but response error is known to be present. We assume that the response standard error is equal to \( \sigma^2 = 225 \) (mg/dl)\(^2\) based on more careful reliability studies using similar measurement protocols (Hegsted and Nicolosi, 1987)). Subtracting this variance component from the residual variance estimated by restricted maximum likelihood using the Seasons’ data, we estimate \( \sigma^2_c \) by 68 (mg/dl)\(^2\) and \( \sigma^2 \) by 1463 (mg/dl)\(^2\). The mean serum cholesterol is given by 218 mg/dl. (Source: sne02p7.sas). Cholesterol screening levels are typically set to identify subjects with serum cholesterol levels that exceed 240 mg/dl. For a realized subject whose observed average serum cholesterol is 240 mg/dl, the predicted latent cholesterol is 238.95 (se=8.20). The predictor of the realized subject’s average cholesterol (including average response error) is 238.96 (se=8.31). (Source: ced02p05.sas). If cholesterol were to be predicted for a 1 week period (\( M = 7 \) days) based on the same data, for a realized subject whose observed average serum cholesterol is 240 mg/dl, the predicted latent cholesterol is 239.1 (se=2.92). The predictor of the realized subject’s average cholesterol (including average response error) is 239.5 (se=5.55). (Source: ced02p06.sas).

5. DISCUSSION

A popular way interpreting realized random effects is via sampling from a population of subjects. The latent values for the subjects in the population are of interest. We observe some response on some subjects and which to predict, for example, a selected subject’s true latent value. In a Bayesian framework, some distributional assumptions are placed on the subject’s latent values (the priors), which may be thought of as a measure of the investigator’s beliefs. Combining the prior information with the observed data results in the predicted posterior distribution. The combination is often based on the assumption that the observed data resulted from a simple random sample from the prior population. The stumbling block in such analyses is often choice of the
prior distribution. Model based approaches in survey sampling refer to the prior distribution as a super-population.

The development that we present in sections two and three is closely related to this Bayesian framework, but removes the stumbling block. Essentially, the prior distribution requires conceptualizing a finite population as a sequence of random variables, as opposed to fixed values. We construct this distribution by assuming permutations of the values in the population are equally likely. The advantage of such a prior model is that we can physically select the observed data to guarantee that the model holds. Strategies for selection are the topic of survey sampling. Thus, the prior distribution corresponds to the sampling model that resulted in the observed data. This gives the methods face validity. It also unifies prediction based approaches to traditional frequentist based survey sampling. It is the philosophical integrity of this approach that makes these methods appealing.

The predictors developed in section three can be compared directly with mixed model or super-population model predictors. In these comparisons, we consider the prediction based approach, as argued by Valliant, Dorfman, and Royall (2000), to have a strong conceptual appeal. For our limited problems, it boils statistical inference down to predicting un-observed values. The best linear unbiased predictor in mixed models (1.4) is the limit of Scott and Smith’s predictor in a super-population model when the fraction of SSUs observed on a realized unit is so small that it can basically be neglected. The traditional BLUP places all the weight on predicting un-observed SSUs. Ignoring the proportion of the SSUs observed has the disadvantage of heightened model sensitivity due to reliance solely on predictors based on the model.

In the balanced two-stage sampling setting considered here, the predictor developed by Scott and Smith (1.5) is nearly identical to the predictor based on the finite population sampling model (3.6). The difference in the predictors results from different definitions of variance components. The two-stage permutation model permutes PSUs, and hence has a single PSU component of variance representing the average SSU variance (given by (2.2)) as opposed to the cluster specific components, $\sigma_i^2$. This, along with the use of a between PSU component of variance defined by (2.21) as opposed to $\delta^2$, results in the difference in predictors. The context described by Scott and Smith appears to match the context for the two-stage sampling, but the model assumptions for the variance components in Scott and Smith’s super-population do not correspond to the two-stage sampling variance components. The model for the super-population is in this sense artificial. Apart from these differences, however, the predictors are the same.

A larger distinction in predictors occurs when response error is present. In such settings, different target random variables can be specified corresponding to the latent value of a PSU (2.17), or the average response for a PSU (2.18). The latent value of a PSU may commonly be the parameter of interest. However, it is possible that there is interest in the average response for a PSU. This parameter arises from the basic definition of the population as a realization from a super-population in model-based survey sampling. For example, the yield of wheat in a given year on a farm (PSU) may be of more interest than the average yield of wheat over several years (the latent value) for the farm.

The models and methods that we present are limited to balanced settings with a single measure for each realized SSU. Extensions to settings with different numbers of
measures on SSUs, to unbalanced situations, and to settings including other covariables are desirable. Settings where there are different numbers of measures on SSUs may be accommodated by taking an average over the measures for an SSU, and defining the response error standard deviation by the response error standard error for the SSU. The interpretation of the average response error variance, (2.1), will differ. Note that this extension assumes that the difference in number of measures per SSU is defined as part of the sampling design.

Extensions to settings where there are different cluster sizes, and possibility different sampling fractions for different clusters are more complicated. We have developed such extensions in a context where there is no response error (see Stanek and Singer, 2002). The additional complication occurs since the number of SSUs for a target parameter must be accounted for when defining the random parameter. Different approaches to simplify this problem have been considered based on collapsing the sample and remainder to totals, or averages for each PSU. Finally, extensions to settings involving covariables either on PSUs and/or SSUs is of obvious interest. Such extensions have been developed in a related context with domains, but not extended to a two-stage sampling setting (see Li, (2002)).

A final caution is warranted for the interpretation of predictors of random variables. An appealing interpretation of the predictor of a realized random effect is that it predicts the latent value for the realized PSU. Thus, if \( U_{ix} = 1 \), the target parameter given by (2.17) is \( \mu_x \). This interpretation, while seemingly self-evident, is not consistent with the development of the predictor. The difference is due to the distinction between the expected value of response for a PSU (at a position), and the expected value of response for a cluster. Aspects of this difference are discussed in the context of simple random sampling by Stanek, Singer, and Lencina (2002). The basic problem can be related to the conventional representation of random variables arising from sampling. In this regard, the work of Godambe (1955) provides a more general framework that may enable a more straightforward strategy for prediction and interpretation.
APPENDIX A. THE VARIANCE UNDER A TWO STAGE RANDOM
PERMUTATION MODEL WITH RESPONSE ERROR

We evaluate the variance of $Y^*$ using the conditional expansion of the variance
given by

\[
\text{var} \left( \frac{Y}{\xi_1, \xi_2} \right) = E_{\xi_1} \left[ \text{var} \left( \frac{Y}{\xi_1, \xi_2} \right) \right] + \text{var} \left( \frac{E_{\xi_1} \left( Y \right)}{\xi_1, \xi_2} \right).
\]

Given $\xi_1$ and $\xi_2$, and using (2.10) and (2.11), since $Y$ is non-stochastic,

\[
\text{var} \left( \frac{Y}{\xi_1, \xi_2} \right) = \left( \frac{\partial^2}{\partial \xi_1 \partial \xi_2} \right) \left[ E_{\xi_1} \left( Y \right) \right] - \frac{\partial}{\partial \xi_1} \left( \frac{\partial}{\partial \xi_2} \right) \left[ E_{\xi_1} \left( Y \right) \right] + \frac{\partial}{\partial \xi_2} \left( \frac{\partial}{\partial \xi_1} \right) \left[ E_{\xi_1} \left( Y \right) \right] - \frac{\partial^2}{\partial \xi_1 \partial \xi_2} \left[ E_{\xi_1} \left( Y \right) \right].
\]

Given $\xi_1$ and $\xi_2$, and using (2.10) and (2.11), since $Y$ is non-stochastic,

\[
\text{var} \left( \frac{Y}{\xi_1, \xi_2} \right) = \left( \frac{\partial^2}{\partial \xi_1 \partial \xi_2} \right) \left[ E_{\xi_1} \left( Y \right) \right] - \frac{\partial}{\partial \xi_1} \left( \frac{\partial}{\partial \xi_2} \right) \left[ E_{\xi_1} \left( Y \right) \right] + \frac{\partial}{\partial \xi_2} \left( \frac{\partial}{\partial \xi_1} \right) \left[ E_{\xi_1} \left( Y \right) \right] - \frac{\partial^2}{\partial \xi_1 \partial \xi_2} \left[ E_{\xi_1} \left( Y \right) \right].
\]

Thus,

\[
\text{var} \left( \frac{Y}{\xi_1, \xi_2} \right) = \left( \frac{\partial^2}{\partial \xi_1 \partial \xi_2} \right) \left[ E_{\xi_1} \left( Y \right) \right] - \frac{\partial}{\partial \xi_1} \left( \frac{\partial}{\partial \xi_2} \right) \left[ E_{\xi_1} \left( Y \right) \right] + \frac{\partial}{\partial \xi_2} \left( \frac{\partial}{\partial \xi_1} \right) \left[ E_{\xi_1} \left( Y \right) \right] - \frac{\partial^2}{\partial \xi_1 \partial \xi_2} \left[ E_{\xi_1} \left( Y \right) \right].
\]

As a result, defining $\sigma^2 = \sum_{i=1}^{M} \sigma_{i,i}^2$, the expectation is given by

\[
E_{\xi_1} \left( \frac{Y}{\xi_1, \xi_2} \right) = \sigma^2 I_M.
\]

Thus,

\[
E_{\xi_1} \left( \frac{Y}{\xi_1, \xi_2} \right) = \sigma^2 I_M.
\]
\[
E_{\xi_{i_{2}}} \left[ \left( U \otimes I_{M} \right) \left( \bigoplus_{i=1}^{N} U^{(i)} \right) \right] D_{\xi_{i_{2}}} \left[ \left( U \otimes I_{M} \right) \left( \bigoplus_{i=1}^{N} U^{(i)} \right) \right]' = \sigma_{\xi_{i_{2}}}^{2} I_{NM}. \quad (A.2)
\]

We evaluate \( \text{var}_{\xi_{i_{2}}} \left( Y \right) = \text{var}_{\xi_{i_{2}}} \left[ E_{\xi_{i_{2}}} \left( Y \right) \right] + \text{var}_{\xi_{i_{2}}} \left[ \text{var}_{\xi_{i_{2}}} \left( Y \right) \right] \) next. The first term in this expansion is given by \( \text{var}_{\xi_{i_{2}}} \left[ E_{\xi_{i_{2}}} \left( Y \right) \right] = \text{var}_{\xi_{i_{2}}} \left( X\mu + ZB \right) \). Now

\[
\text{var}_{\xi_{i_{2}}} \left[ E_{\xi_{i_{2}}} \left( Y \right) \right] = \text{var}_{\xi_{i_{2}}} \left( U\beta \right) \otimes J_{M}, \quad \text{where} \quad \text{var}_{\xi_{i_{2}}} \left( U\beta \right) \text{ is equal to the variance of a random permutation of PSU means in the population of PSUs. As a result,}
\]

\[
\text{var}_{\xi_{i_{2}}} \left[ E_{\xi_{i_{2}}} \left( Y \right) \right] = \sigma_{\xi_{i_{2}}}^{2} \left( P_{N} \otimes J_{M} \right), \quad \text{where} \quad P_{a} = I_{a} - \frac{1}{a} J_{a}. \quad \text{We evaluate the second term in}
\]

the conditional expansion of the variance in a similar manner. For the cluster in position \( i^{th} \) in a given permutation of clusters, the variance of the SSUs is \( \text{var}_{\xi_{i_{2}}} \left( y_{i} \right) = s_{i}^{2} P_{M} \)

where \( s_{i}^{2} \) is the realization of the random variable \( S_{i}^{2} = \sum_{a=1}^{N} U_{a} \sigma_{s_{i}}^{2} \). Also, due to independence of the permutations of SSUs, \( \text{cov}_{\xi_{i_{2}}} \left( Y_{i}, Y_{j} \right) = 0 \) for \( j \neq j^{*} \). Combining these results for all \( i = 1, ..., N \), \( \text{var}_{\xi_{i_{2}}} \left( Y_{i} \right) = D_{s} \otimes P_{M} \) where \( D_{s} \) is a diagonal matrix with diagonal elements equal to the elements in the vector \( s = \left( \left( s_{i}^{2} \right) \right) \). Taking the expectation over random permutations of SSUs, \( E_{\xi_{i_{2}}} \left( S_{i}^{2} \right) = \sum_{i=1}^{N} \frac{\sigma_{s_{i}}^{2}}{N} = \sigma_{s_{i}}^{2} \), and hence

\[
\text{var}_{\xi_{i_{2}}} \left( Y \right) = \sigma_{\xi_{i_{2}}}^{2} \left( P_{N} \otimes J_{M} \right) + \sigma_{s_{i}}^{2} \left( I_{N} \otimes P_{M} \right). \quad (A.3)
\]

Adding (A.2) to(A.3) results in (2.20).
References


