Predicting Realized Cluster Parameters from Two Stage Samples of Unequal Size Clustered Population

Edward J. Stanek III
Department of Biostatistics and Epidemiology
401 Arnold House
University of Massachusetts
715 North Pleasant Street
Amherst, MA 01003-9304 USA
stanek@schoolph.umass.edu

Julio Singer
Departamento de Estatística
Universidade de São Paulo
São Paulo, Brazil
jmsinger@ime.usp.br
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: super-population, best linear unbiased predictor, two-stage sampling, random permutation, optimal estimation, design-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 Loureiro. 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

We develop design-based methods for predicting the mean or total of a realized cluster based on a two-stage cluster random sample from a finite population formed from unequal size clusters. Prediction of random effects is commonly based on mixed model theory (Goldberger 1962; Henderson 1984; McLean, Sanders and Stroup 1991; Robinson 1991; McCulloch and Searle 2001) leading to best linear unbiased predictors (BLUP). The mixed model assumes an infinite population of clusters and units, and hence does not account for differences in cluster sizes. In the survey sampling literature, predictors of the mean of realized clusters have been developed using super-population models (Scott and Smith 1969; Bolfarine and Zacks 1992; Valliant, Dorfman and Royall 2000) that account for different size clusters of super-population random variables. Design-based predictors based on a two-stage random permutation model have been developed, but are limited to equal cluster size settings (Stanek and Singer 2004), possibly including response error.

We extend the random permutation framework to clustered populations where the number of units in clusters are not equal, and use it to predict target random variables that correspond to the average (or total) response for units in a cluster (which we call the latent value). The prediction is based solely on the two stage sampling design; no hypothetical super-population is required, and accounting is made for different size clusters in the finite population.

Mixed model and super-population model approaches can be directly applied to estimate fixed effects in unequal size clustered populations. These approaches can be contrasted with design-based approaches in survey sampling (Sarndal, 1992 #1024) (Cochran 1977) where estimators depend on sampling fractions, and become more complex along with the sampling design. The additional complexity is due in part to maintaining an account of the population complexity.

Predicting random effects based on a two-stage random permutation model for cluster sampling is complicated by unequal size clusters that necessitates careful tracking of target random variables through the development of predictors. Tracing through this development, while more complex, retains a clear connection between the target random variable, and the predictor. Avoiding these complications with other methods comes with a cost- more extensive assumptions and ambiguity in the interpretation of predictors.

As an example, we consider a cluster randomized controlled trial designed to evaluate the relative effectiveness of a training program for physician-delivered nutrition counseling, alone and in combination with an office support program, on dietary fat intake and blood LDL-cholesterol levels in patients with hyperlipidemia (the Worcester-Area Trial for Counseling in Hyperlipidemia (WATCH) Trial) (Ockene, Hebert, Ockene, Saperia, Nicolosi, Merriam and Hurley 1996). Forty-five primary care internists were selected from 60 internists at the Fallon Clinic, a central Massachusetts health maintenance organization (HMO), with each internist randomly assigned to one of three conditions: (I) Usual Care; (II) Physician nutrition counseling training; and (III) Physician nutrition counseling training plus an office support program. Twelve hundred and seventy-eight patients with blood LDL-cholesterol levels in the highest 25th percentile, having previously-scheduled physician visits, were recruited into the study. The number of sampled patients per cluster ranged from 1 to 43.
The WATCH study had a cluster randomized design, with patients in physician practices forming the basic clusters. We focus our attention on a single dependent variable, percent of Kcal from saturated fat, which we consider as a difference constructed by taking the baseline minus one year measure for each patient. There was considerable variability in the magnitude of the average patient’s saturated fat change between physician’s practices within a given intervention. We discuss estimating the change for eligible patients (units) in a physician’s practice (cluster) that accounts for the practice size.

1.1 Basic Notation

We define basic notation and terminology for a clustered finite population, and then use it to provide an overview of prediction based on mixed models (Searle, Casella and McCulloch 1992; Verbeke and Molenberghs 2000; McCulloch and Searle 2001; Diggle, Heagerty, Liang and Zeger 2002) and super-population sampling models (Bolfarine and Zacks 1992; Valliant, Dorfman et al. 2000). Let a finite population be defined by a listing of $t = 1, ..., M_s$ units in each of $s = 1, ..., N$ clusters, where response for unit $t$ in cluster $s$ is given by $y_{st}$. We assume that a two-stage cluster sample of $n$ clusters from this population (without replacement) is to be selected, with a different (but known) number of units, $m_s$, selected from each of the selected clusters.

We represent the sample of clusters by defining a random permutation of the clusters in the original frame. A new label, $i = 1, ..., N$, is assigned to the positions in a permutation of clusters. Without loss of generality, we assume that the first $i = 1, ..., n$ positions contain the sample clusters. In a similar manner, we distinguish the listed units in cluster $s$ from a random permutation of units in cluster $s$, whose positions are labeled by $j = 1, ..., M_s$. We again assume that the first $j = 1, ..., m_s$ positions contain the units in the sample if cluster $s$ is selected. 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 unit that will occupy position $j$ in the permutation as the $j^{th}$ secondary sampling unit (SSU). Using these ideas and notation, we represent a two-stage permutation of the population as an ordered list of $N = \sum_{s=1}^{N} M_s$ random variables, where the random variable representing response for SSU $j$ in PSU $i$ is $Y_{ij}$. PSUs and SSUs are indexed by positions ($i$ and $j$), whereas clusters and units are indexed by labels ($s$ and $t$) in the finite population.

We think of sampling as selecting a permutation. 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 position $i$ is a random variable. Once the sample has been selected, for positions $i \leq n$, we will observe which cluster corresponds to a particular PSU in the sample. We refer to that cluster as the realized PSU, and to the average (or total) response for units in that cluster as the latent value of the realized PSU (noting that such values are often defined as an overall mean plus a deviations relative to the overall mean). In this context, mixed model and super-population modeling strategies may be advocated for predicting the latent value of a realized PSU, or linear combinations of such latent values. We first review these
approaches. We then define the context for finite population sampling prediction 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 SSU $j$ in PSU $i$ is given by

$$ Y_{ij} = \mu + B_i + E_{ij} \quad (1) $$

where $\mu$ corresponds to the overall population mean, $B_i$ is a random effect that corresponds to the deviation from the population mean of the latent value of PSU $i$, and $E_{ij}$ is the random error corresponding to the deviation of response for SSU $j$ from the mean of PSU $i$.

Typically, additional assumptions are made. The random effects, $B_i$, are assumed independent and identically normally distributed, $B_i \sim iid \ N(0, \sigma^2_i)$, and independent of the random errors, also assumed independent and identically distributed for all $j$ in $i$, $E_{ij} \sim iid \ N(0, \sigma^2_j)$. 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, Searle and von Krosigk 1959). In model (1), assuming known variances, the estimator of the fixed effect is the weighted least squares estimator, $\hat{\mu} = \sum_{i=1}^{n} w_i \bar{Y}_i$ where

$$ w_i = \frac{1}{\nu_i} = \frac{1}{\sum_{j=1}^{m} \frac{Y_{ij}}{v_i}}, \text{ and } \nu_i = \sigma^2 + \frac{\sigma^2_i}{m_i}. \quad \text{Solution of the mixed model equations results in the BLUP of } B_i. $$

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

$$ \hat{\mu} + k_i \left( \bar{Y}_i - \hat{\mu} \right) \quad (2), $$

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

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 (2) 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 that different cluster size has on the predictors.

1.3 Super-population Sampling Models
Predictors of a linear combination of fixed and random effects (corresponding to the average response for a PSU) in two-stage sampling settings were developed by Scott and Smith (1969) using a model based survey sampling approach. 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 it. The parameters of interest are linear combinations of values in the finite population, such as the latent value for a realized PSU. Such parameters can be defined for all clusters in the finite population. Since only a part of the finite population is observed in a sample, the essential statistical problem is how best to predict the values of the remaining random variables. Predictors are constructed using the joint distribution assumed for the super-population (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 the nested super-population by a model, and first and second moments. The random variables $Y_{ij}$, $i = 1,\ldots,N; j = 1,\ldots,M_i$ in the super-population are assumed to satisfy $E(Y_{ij}) = \mu$ and

$$\text{cov}(Y_{ij}, Y_{kl}) = \delta^2 + \sigma_i^2 \quad \text{when } i = k; j = l$$

$$= \delta^2 \quad \text{when } i = k; j \neq l$$

$$= 0 \quad \text{otherwise.}$$

Scott and Smith (1969) used this super-population model to derive predictors of linear combinations of elements in the super-population, and in particular, the linear combination that corresponds to the latent value of a PSU. Scott and Smith frame the statistical problem as prediction of the remaining unobserved SSUs in a 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 assuming that PSUs and SSUs within PSUs are exchangeable) are given, and shown to result in the same predictor given by

$$f_i\bar{Y} + (1 - f_i)
\left[\hat{\mu}^* + k_i^* \left(\bar{Y}_i - \hat{\mu}^*\right)\right] (3)$$

if the PSU is partially realized in the sample, or $\hat{\mu}^*$ if the PSU is not included in the sample, where

$$\hat{\mu}^* = \sum_{i=1}^n w_i^* \bar{Y}_i, \quad f_i = \frac{m_i}{M_i}, \quad w_i^* = \frac{1}{1/v_i^*}, \quad v_i^* = \delta^2 + \frac{\sigma_i^2}{m_i} \quad \text{and} \quad k_i^* = \frac{\delta^2}{v_i^*}.$$  

This predictor has an appealing interpretation as the weighted sum of two terms: the sample mean for the realized PSU, and predictors of the remaining SSUs for the realized PSU. The weighting factors are the proportion of SSUs that are observed and the proportion that are not observed. For PSUs that are not in the sample, the predictor simplifies to the weighted sample mean.

There is an obvious similarity between equation (2) and equation (3). When $f_i$ is small enough that the first term in equation (3) can be ignored, and $1 - f_i \approx 1$, the two equations appear to be identical, with the exception of different representations for the variance components. The best linear unbiased predictor (in equation (2)) predicts 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 sampled, and predicted values of the unobserved SSUs. 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. Instead, the model assumptions correspond to assuming 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.

2. THE TWO-STAGE PERMUTATION MODEL FOR A FINITE UNEQUAL SIZE CLUSTERED POPULATION

We derive predictors of realized random effects from a two-stage sample of a finite unequal size clustered population using a random permutation model similar to {Stanek, 2004 #1697}. A principal distinction is the use of a probability model based directly on the two-stage sampling and thus is based on the sample design. The results, while similar in form to those of Scott and Smith 1969 and Valliant, Dorfman et al. 2000, do not require positing a super-population. We first define the population and target parameters. Next we define the random variables that represent a two-stage random permutation of the population, and variations in the representation that preserve representation of the cluster’s identity.

2.1 Finite Population Parameters and Re-parameterizations

We begin by defining finite population parameters and representing response for a subject in a simple deterministic model. The mean and variance of cluster $s$ are defined as

$$\mu_s = \frac{1}{M_s} \sum_{t=1}^{M_s} y_{st} \quad \text{and} \quad \sigma^2_s = \frac{1}{M_s} \sum_{t=1}^{M_s} (y_{st} - \mu_s)^2$$

for $s = 1, ..., N$ (where we use the survey sampling definition of the parameter $\sigma^2_s$). Similarly, the mean of clusters in the population, and the variance between cluster means are 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

$$y_{st} = \mu + \beta_s + \varepsilon_{st} \quad (4)$$

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 (4) is called a derived model (Hinkelmann and Kempthorne 1994).
Defining $y = \begin{pmatrix} y_1' & y_2' & \cdots & y_N' \end{pmatrix}'$ where $y_s = \begin{pmatrix} y_{s1} & y_{s2} & \cdots & y_{sM_s} \end{pmatrix}'$, model (4) can be summarized as

$$y = X\mu + Z\beta + \varepsilon \quad (5)$$

where $X = I_n$, $Z = \sum_{s=1}^{N} I_{M_s}$ and $\beta' = (\beta_1 \beta_2 \cdots \beta_N)$. Here, $I_s$ is an $a \times 1$ column vector of ones, and $\bigoplus A_s$ denotes a block diagonal matrix with blocks given by $A_s$ (Graybill 1983), and $\varepsilon$ is defined similarly to $y$. None of the terms in model (5) are random variables.

2.2 Random Variables and The Two Stage Random Permutation Model

We define a stochastic model that parallels model (5) for two-stage cluster sampling. The sample of clusters corresponds to the first $n$ clusters in a permutation of the vectors of clusters, $y_s$. Similarly, the sample of units in cluster $s$ corresponds to the first $m_s$ units in a permutation of cluster units. We 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).

For each permutation, we assign a new label, $i = 1, \ldots, N$, to the clusters according to its position in the permuted list. In a similar manner, we assign a new label, $j = 1, \ldots, M_s$, to the units according to its position in the permuted list for cluster $s$. We refer to the cluster that will occupy position $i$ in the permutation of clusters as primary sampling unit (PSU) $i$, and to the unit that will occupy position $j$ in a permutation of units in a cluster as secondary sampling unit (SSU) $j$. In the finite population, we represent the response for unit $t$ in cluster $s$ by $y_{st}$; in the permuted population, we represent the random variable for response of the unit in position $j$ for the cluster in position $i$ by $Y_{ij}$. PSUs and SSUs are indexed by positions ($i$ and $j$), whereas clusters and units are indexed by labels ($s$ and $t$) in the finite population.

Sample indicator random variables relate $y_{st}$ to $Y_{ij}$. The indicator random variable $U_{st}$ takes on a value of one when PSU $i$ is cluster $s$, and a value of zero otherwise; the indicator random variable $U_{jt}^{(s)}$ takes on a value of one when SSU $j$ in cluster $s$ is unit $t$, and zero otherwise. As a consequence, the random variable corresponding to PSU $i$ and SSU $j$ in a permutation is given by

$$Y_{ij} = \sum_{s=1}^{N} \sum_{t=1}^{M_s} U_{st} U_{jt}^{(s)} y_{st} \quad (6).$$

Defining $U^{(s)} = \begin{pmatrix} U_{1}^{(s)} & U_{2}^{(s)} & \cdots & U_{M_s}^{(s)} \end{pmatrix}'$ where $U_{t}^{(s)} = \begin{pmatrix} U_{1t}^{(s)} & U_{2t}^{(s)} & \cdots & U_{M_st}^{(s)} \end{pmatrix}'$ is an $M_s \times 1$ column vector, a vector representing a permutation of units in a given cluster is given by the $M_s \times 1$ dimension vector $U^{(s)} y_s$. With this notation, the SSUs for the cluster in position $i$ in a permutation of clusters are of the form $U_{it} U^{(s)} y_s$. The dimension of this vector depends on
which cluster occurs in position \( i \) in a permutation, as determined by the realization of the random variables \( U_{s} \), \( s = 1, \ldots, N \).

Since these sizes differ for different clusters, when cluster vectors are permuted, the interpretation of which unit corresponds to SSU \( j \) in PSU \( i \) in a vector representing a two-stage permutation of the population is problematic. For example, if \( N = 3 \), \( M_1 = M_2 = 2 \), and \( M_3 = 3 \), consider two permutations of clusters. Suppose the first permutation has cluster \( s = 1 \) in position \( i = 1 \); \( s = 2 \) in position \( i = 2 \); \( s = 3 \) in position \( i = 3 \). Assume the second permutation has clusters 3, 1, and 2 in positions \( i = 1, \ldots, 3 \), respectively. We can represent the population by a random vector for the first permutation as

\[
\left( \begin{array}{c}
(U_1^{(1)} y_1)'
\end{array}\right) \left( \begin{array}{c}
(U_2^{(2)} y_2)'
\end{array}\right) \left( \begin{array}{c}
(U_3^{(3)} y_3)'
\end{array}\right)'
\]

and for the second permutation as

\[
\left( \begin{array}{c}
(U_3^{(3)} y_3)'
\end{array}\right) \left( \begin{array}{c}
(U_1^{(1)} y_1)'
\end{array}\right) \left( \begin{array}{c}
(U_2^{(2)} y_2)'
\end{array}\right)'
\]

Although both vectors are of dimension \( N \times 1 \) (where \( N = 7 \)), the third SSU in the first permutation is in PSU \( i = 2 \), while the third SSU in the second permutation is in PSU \( i = 1 \). The position of the SSUs in the permuted population is not sufficient to retain the identity of the PSU for the SSU. Basically, the problem is that the dimension of the random variables in a PSU depends on the dimension of the cluster that is realized. This problem is a result of limitations in the traditional representation used for random variables in the context of two stage permutations. The ambiguity is a result of inadequate notation; it gives rise to ambiguity in analysis and interpretation.

The notational ambiguity is avoided by expanding the set of random variables that represent the two-stage permutation to a vector of dimension \( \left( \sum_{s=1}^{N} M_s^2 \right) \times 1 \) whose elements are of the form \( R_{sij}^* = U_{ia} U_{j}^{(s)} y_{st} \). We arrange these random variables in the vector \( R^* \) where

\[
R^* = \left( R_{1+}^* \right) \left( R_{2+}^* \right) \cdots \left( R_{N+}^* \right)'
\]

with \( R_{s+}^* = \left( \begin{array}{c}
R_{s1}^* \\
R_{s2}^* \\
\vdots \\
R_{sN}^*
\end{array}\right)' \) and where

\[
R_{s+}^* = \left( U_s \otimes U_i^{(s)} \right)y_{st}
\]

is of dimension \( N M_s \times 1 \) where \( U_s = \left( U_{1s} \ U_{2s} \ \cdots \ U_{Ns} \right)' \) for \( s = 1, \ldots, N \). The position of the random variables in \( R^* \) uniquely identify a cluster and unit. Using the expanded representation and the parameterization from (5), the mixed model is represented by

\[
R^* = X \mu + \left( \begin{array}{cc}
\beta_1 & 0 \\
0 & \beta_2 \\
\vdots & \vdots \\
0 & \beta_N
\end{array}\right) + \left( \begin{array}{cc}
\varepsilon_1 & 0 \\
0 & \varepsilon_2 \\
\vdots & \vdots \\
0 & \varepsilon_N
\end{array}\right) + \varepsilon (7)
\]

where \( X_s = \frac{1}{N M_s^2} \) and \( X_{s+} = I_{M_s} \otimes \frac{1}{N M_s} \).

Expressions for the expected value and variance of \( R^* \) can be determined using elementary properties of the indicator random variables. Under the two-stage random
permutation model, using the subscript $\xi_1$ to denote expectation with respect to permutations of the clusters and using the subscript $\xi_2$ to denote expectation with respect to permutations of units in a cluster,

$$E_{\xi_1,\xi_2}(\mathbf{R}^*) = \begin{pmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_N \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix} + \begin{pmatrix} \mathbf{X}_{1^+} & 0 & \cdots & 0 \\ 0 & \mathbf{X}_{2^+} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_{N^+} \end{pmatrix} \mathbf{e}, \quad (8)$$

and

$$\text{var}_{\xi_1,\xi_2}(\mathbf{R}^*) = \sum_{s=1}^{N} \left( \frac{1}{M_s - 1} \left( \otimes_{i=1}^{M_s} y_{s_i}^2 - \frac{1}{N} \mathbf{y}_s \mathbf{y}_s' \right) \bigotimes_{s=1}^{M_s} \mathbf{P}_M \right) \mathbf{1}_N \bigotimes \mathbf{P}_N + \left( \frac{1}{N-1} \mathbf{y}_s \mathbf{y}_s' \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{J}_{M_s}}{M_s} \right)$$

$$- \frac{1}{N(N-1)} \begin{pmatrix} \mathbf{y}_1 \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \\ \mathbf{y}_2 \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \\ \vdots \\ \mathbf{y}_N \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \end{pmatrix} \left( \begin{pmatrix} \mathbf{y}_1 \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \\ \mathbf{y}_2 \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \\ \vdots \\ \mathbf{y}_N \bigotimes \mathbf{P}_N \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \end{pmatrix} \right)' \quad (9)$$

where $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a} \mathbf{J}_a$.

2.3. Defining Random Variables of Interest

We assume that there is interest in a linear combination of PSU totals, $P^* = \mathbf{g}'' \mathbf{R}^*$ or means defined by $P^* = \mathbf{g}'' \mathbf{R}^*$ where $\mathbf{g}'' = \mathbf{1}_N' \left( \bigotimes_{s=1}^{N} \mathbf{1}_{M_s} \bigotimes \left[ \mathbf{b} \bigotimes \mathbf{1}_{M_s} \right] \right)$ (for totals) or

$$\mathbf{g}'' = \mathbf{1}_N' \left( \bigotimes_{s=1}^{N} \mathbf{1}_{M_s} \bigotimes \left[ \mathbf{b} \bigotimes \frac{\mathbf{1}_{M_s}}{M_s} \right] \right)$$

(for means), where $\mathbf{b} = (b_i)$ is an $N$-dimensional column vector of constants. Of principal interest is the linear combination that defines the total for PSU $i$ given by $P_i^* = \mathbf{g}_i'' \mathbf{R}^*$, or the mean for PSU $i$ given by $P_i^* = \mathbf{g}_i'' \mathbf{R}^*$, where

$$\mathbf{g}_i'' = \mathbf{1}_N' \left( \bigotimes_{s=1}^{N} \mathbf{1}_{M_s} \bigotimes \left[ \mathbf{e}_i \bigotimes \mathbf{1}_{M_s} \right] \right) \quad (10)$$

and
\[ g_{is}^* = \mathbf{1}_{N}^t \left( \sum_{i=1}^{N} \mathbf{1}_{M_i}' \otimes \left( \mathbf{e}_i' \otimes \mathbf{1}_{M_s}^t \right) \right) , \]  \hspace{1cm} (11)

where \( \mathbf{e}_i \) is an \( N \)-dimensional column vector with the \( i^{th} \) value equal to one, and other values equal to zero. Note that

\[ P_i^* = \sum_{s=1}^{N} U_{is} M_s \mu_s , \]  \hspace{1cm} (12)

while

\[ P_i^* = \sum_{s=1}^{N} U_{is} \mu_s . \]  \hspace{1cm} (13)

We develop predictors of the random variables given by (12) and (13) next.

3. PREDICTING A PSU TOTAL OR MEAN BASED ON A TWO-STAGE SAMPLE

We develop a predictor of a PSU total (12) (or mean (13)) based on the two-stage without replacement simple random sampling probability model (7), where sampling results in a single non-stochastic measure on each of \( m_s \) SSUs from each of the \( n \) PSUs. We require the predictor to be a linear function of the sample, to be unbiased, and to minimize 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). The elements of \( \mathbf{R}^* \) are first partitioned into a sampled and remaining portion. We assume that the elements in the sampled portion will be observed, and express \( P_i^* \) as the sum of two parts, one which is a function of the sample, and the other which is a function of the remaining random variables. Then, requiring the predictor to be a linear function of the sampled random variables and to be unbiased, coefficients are evaluated that minimize the
expected value of the mean squared error, $\text{var}_{\hat{\xi}}\left(\hat{P}_i - P_i\right)$. This general development was given in Theorem 2.1 by (Royall 1976).

TO HERE 2/12/03

Conceptually, Royall’ theorem can be applied directly to form predictors, but practically is complicated by singularities in $\text{var}\left(R^*\right)$. The singularities are the result of the expansion of random variables needed to explicitly track clusters. We avoid these complications by developing solutions to a simpler problem based on a ‘collapsed’ set of random variables, $Y$. To do so, we first express $R^* = A'Y + B'R^*$. Different definitions of $Y$ are possible. For definitions of $Y$ that result in $P_i = g_i''A'Y$ and $B'R^* = 0$, we develop predictors of $P_i$ based solely on $Y$. It turns out that in this setting, unbiased predictors can be developed only when the sampling fraction is constant for all clusters. In settings where the sampling fraction differs between clusters, we develop an unbiased predictor of $P_i^* = g_i''Y^*$, where $P_i^* = E_{\hat{\xi}}\left(P_i^*\right)$ is given by (13).

The key step in these developments is an appropriate definition of the ‘collapsed’ set of random variables, $Y$. We consider three different definitions of $Y$. The first definition, $Y^* = C''R^*$, collapses random variables to PSU totals for the sample and remainder. The second definition, $Y^* = C''R^*$, collapses random variables to scaled PSU totals for the sample and remainder where the scaling factor is the cluster size. A third strategy, $Y^* = C''R^*$, is to collapse random variables to PSU means for the sampled and remaining units. We proceed to outline details of each of these developments.

3.1. Predictors of PSU Totals Based on Collapsing to Cluster Totals
We develop the best linear unbiased predictor of (12) based on collapsing $\mathbf{R}^*$ to a vector of PSU totals for the sample and remainder such that $\mathbf{Y}^* = \mathbf{C}^* \mathbf{R}^*$ where

$$\mathbf{C}^* = \begin{pmatrix} C_1^* & C_2^* & \cdots & C_N^* \end{pmatrix}$$

and

$$\mathbf{C}^*_{2N \times NM^2} = \left( \mathbf{1}_{M_s}^t \otimes \mathbf{I}_N \right) \otimes \begin{pmatrix} \mathbf{1}_{m_s}^t & \mathbf{0} & \mathbf{1}_{m_s}^t \end{pmatrix}_{1 \times (M_s - m_s)} \quad (1).$$

For PSU $i$, the corresponding elements of $\mathbf{Y}^*$ are given by

$$\overline{y}_{si} = \frac{1}{m_s} \sum_{j=1}^{m_s} \left( \sum_{i=1}^{M_s} U_{j(i)} y_{si} \right) \quad \text{and} \quad \overline{y}_{sii} = \frac{1}{M_s - m_s} \sum_{j=m_s+1}^{M_s} \left( \sum_{i=1}^{m_s} U_{j(i)} y_{si} \right).$$

Defining

$$\mathbf{A}' = \left( \bigoplus_{i=1}^N \mathbf{C}_i \left( \mathbf{C}_i \mathbf{C}_i^* \right)^{-1} \right) \left( \frac{\mathbf{I}_N}{N} \otimes \mathbf{I}_{2N} \right),$$

$$\mathbf{B}' = \left( \bigoplus_{i=1}^N \mathbf{C}_i \left( \mathbf{C}_i \mathbf{C}_i^* \right)^{-1} \right) \left( \mathbf{P}_{N \times I_{2N}} \otimes \mathbf{I}_{2N} \right) + \mathbf{J}_{M_s} \otimes \mathbf{I}_N \times \mathbf{P}_{M_s \times I_{NM},} \quad (2)$$

we can express $\mathbf{R}^* = \mathbf{A}' \mathbf{Y}^* + \mathbf{B}' \mathbf{R}^*$ (see c02ed34.doc). Then using (10), we find that $\mathbf{P}^* = \mathbf{g}' \mathbf{R}^* = \mathbf{g}' \mathbf{Y}^*$.

Also, $\mathrm{var}_{\mathbf{Y}^*} \left( \mathbf{g}' \mathbf{R}^* \right) = \mathrm{var}_{\mathbf{Y}^*} \left( \mathbf{g}' \mathbf{Y}^* \right)$ (see c02ed37.doc, p1). This provides the rational for developing predictors of (2) based on $\mathbf{Y}^*$.

We highlight the main results for the development. More details of the development are given by Stanek (2002, c02ed11, 12v1, 34-38.doc). Using $\mathbf{C}_i$ defined by (1), model (7) simplifies to

$$\mathbf{Y}^* = \left( \mathbf{1}_N \otimes \mathbf{I}_2 \right) \left( \begin{pmatrix} \overline{\tau}_f \overline{\tau}_f \end{pmatrix} + \mathbf{I}_{2N} \begin{pmatrix} T_f \mathbf{T}_f \end{pmatrix} \right) + \mathbf{E}^* \quad (4)$$

where

$$T_f = \sum_{s=1}^N U_{is} \left( f_s \tau_s - \overline{\tau}_f \right), \quad T_i = \left( \sum_{s=1}^N U_{is} \left( \tau_s - \overline{\tau} \right) \right), \quad \overline{\tau}_f = \frac{\sum_{s=1}^N f_s \tau_s}{N}, \quad \overline{\tau} = \frac{\sum_{s=1}^N \tau_s}{N},$$

$$\tau_s = M_s \mu_s, \quad f_s = \frac{m_s}{M_s}, \quad \text{and} \quad \mathbf{E}^* = \mathbf{C}^* \mathbf{E}.$$ Model (4) is a mixed model where the expected value and variance of $\mathbf{Y}^*$ are given by
\[ E_{st} \right( Y^* \right) = \left( I_N \otimes I_2 \right) \left( \frac{\bar{f}_i}{\bar{f}_i - \bar{f}_j} \right) \]  

and

\[ \text{var}_{st} \right( Y^* \right) = I_N \otimes \left( \frac{\sum_{s=1}^{N} \frac{M_s \sigma_s^2}{N}}{N} \right) \left( \frac{f_s}{1-f_s} \right) \right) + \left( \frac{1}{N-1} \right) \left( I_N - \frac{J_N}{N} \right) \otimes \left( \frac{\sum_{s=1}^{N} \left( f_s \right)^2 - \left( f_s \right)^2 \left( 1-f_s \right) \left( 1-f_s \right) \right) \right) \right) - N \left( \frac{\bar{f}_i}{\bar{f}_i - \bar{f}_j} \right) \left( \frac{\bar{f}_i}{\bar{f}_i - \bar{f}_j} \right) \right) \].  

The model contains two types of random effects. One random effect, \( T_i \), represents deviations of cluster totals about the simple average cluster total for the \( i^{th} \) selected cluster. The second random effect, \( T_{II} \), represents deviations of the expected cluster sample totals about the simple average expected cluster sample total for the \( i^{th} \) selected cluster.

We develop a predictor of (2) corresponding to the total of the \( i^{th} \) PSU in the permutation. We require the predictor to be a linear function of the sampled random variables, to be unbiased, and to minimize the expected value of the mean squared error. First, we re-arrange \( Y^* \) into a sampled and remaining vector by pre-multiplying by

\[ K^* = \begin{pmatrix} K_t^* \\ K_{II}^* \end{pmatrix} \]

where

\[ K_t^* = \begin{pmatrix} I_n \mid 0 \end{pmatrix}_{n \times (N-n)} \otimes \begin{pmatrix} 1 \mid 0 \end{pmatrix} \]  

and

\[ K_{II}^* = \begin{pmatrix} \begin{pmatrix} I_n \mid 0 \end{pmatrix}_{n \times (N-n)} \otimes \begin{pmatrix} 0 \mid 1 \end{pmatrix} \end{pmatrix}_{(N-n) \times N} \]  

resulting in \( K^* Y^* = \begin{pmatrix} Y_t^* \\ Y_{II}^* \end{pmatrix} \). In a similar manner, we define \( K^* X = K^* [I_N \otimes I_2] \) where

\[ X_t = I_n \otimes \begin{pmatrix} 1 \mid 0 \end{pmatrix} \]  

and \( X_{II} = \begin{pmatrix} I_n \otimes \begin{pmatrix} 0 \mid 1 \end{pmatrix} \end{pmatrix}_{N \times (N-n)} \otimes I_2 \). In terms of the partitioned random variables,

\[ P^*_i = g_{it} Y_t^* + g_{it} Y_{II}^* \]  

where \( g_{it} = e_{it} \) and \( g_{it} = \left( e_{it} \otimes I_2 \right) \), and \( e_{it} = \left( e_{it} \otimes I_2 \right) \) where \( e_{it} \) is an \( n \times 1 \) vector, and \( e_{it} \) has dimension \( (N-n) \times 1 \).

Collection of the study data will result in realizing the values of \( Y^*_t \). We require the predictor of \( P^*_i \) to be a linear function of the sample data, \( \hat{P}^*_i = (g_{it} + a^*) Y_t^* \) and be unbiased.
such that $E_{\xi_i \xi_i} \left( \hat{P}_i - P_i \right) = 0$. Since $E_{\xi_i \xi_i} \left( Y_i - \bar{\tau} \right) = X_i \left( \bar{\tau}_f \right)$ and $E_{\xi_i \xi_i} \left( Y_{ii} \right) = X_{ii} \left( \bar{\tau}_f \right)$, the unbiased constraint implies that \( a'X_i - g_{ii}X_{ii} \left( \bar{\tau}_f \right) = 0 \), or that $a'X_i = g_{ii}X_{ii}$.

This expression simplifies to

\[
\left( a'1_n \mid 0 \right) = \left( e_{ii} 1_{N-n} \mid 1 \right) \tag{9}
\]

which is clearly impossible, since 0 cannot equal 1. As a result, unbiased predictors of (2) are not possible for model (4).

Suppose, however, that second stage samples are selected with probability proportional to size, such that $f = \frac{m_s}{M_s}$ for all $s = 1, \ldots, N$. Then $\bar{\tau}_f = f \bar{\tau}$, and the unbiased constraint can be expressed as $a'X_i - g_{ii}X_{ii} \left( \bar{\tau}_f \right) = 0$ implying that

\[
\left( a'X_i - g_{ii}X_{ii} \right) \left( f 1_f \right) = 0. \tag{10}
\]

The unbiased constraint now simplifies to

\[
a'1_n = e'_{ii} 1_{N-n} + \frac{1-f}{f} \]. These conditions will be satisfied if $a'1_n = \frac{M_s - m_s}{m_s}$ for all $s = 1, \ldots, N$ when $i = 1, \ldots, n$ (corresponding to predicting a PSU in one of the first $n$ positions in the permutation), or if $a'1_n = \frac{M_s}{m_s}$ for all $s = 1, \ldots, N$ when $i = n+1, \ldots, N$ (when predicting a PSU in one of the last $n+1, \ldots, N$ positions in the permutation (see c02ed11.doc)).

We assume second stage samples are selected with probability proportional to size in the subsequent developments, and proceed to develop a predictor based on minimizing the expected value of the mean squared error. Using this notation, model (4) simplifies to

\[
Y^* = X^* \bar{\tau} + Z^* T + E^* \tag{10}
\]

where $X^* = 1_N \otimes \left( f 1_f \right)$, $Z^* = 1_N \otimes \left( f 1_f \right)$, $\tau = \left( \tau_1 \quad \tau_2 \quad \cdots \quad \tau_N \right)'$, and $T = \mathbf{U} \left( \tau - 1_N \bar{\tau} \right)$. Model (10) is a mixed model for cluster totals. The expression for the variance (6) simplifies to (see c02ed12v1.doc, p8)

\[
\text{var}_{\xi_i \xi_i} \left( Y^* \right) = \sigma^2_{\epsilon_i} 1_N \otimes \left[ \begin{array}{cc}
\left( f 0 \right) & -\left( f 1_f \right) \left( f 1_f \right)' \\
0 & \left( 1_f \right) \left( 1_f \right)'
\end{array} \right] + \sigma^2_{\tau} \left( 1_N - \frac{J_N}{N} \right) \otimes \left( f 1_f \right) \left( f 1_f \right)' \tag{11}\]

\[
+ \sigma^2_{\epsilon_i} 1_N \otimes \left( f 1_f \right) \left( f 1_f \right)'
\]
where \( \sigma^2 = \frac{\sum_{s=1}^{N}(\tau_s - \bar{\tau})^2}{N-1} \), and \( \sigma^2_{e\tau} = \frac{\sum_{s=1}^{N}M_s\sigma^2_{e\tau}}{N} \).

The best linear unbiased predictor of the \( i^{th} \) selected cluster total is given by

\[
\hat{P}_i = g_{i1}^T Y_i + g_{i1}^T \left[ X_{i1}^\tau \hat{\alpha} + V_{i1}^\tau \left( Y_i - X_i^\tau \hat{\alpha} \right) \right] \text{ where } X_i = f1_n, \quad X_i^T = \begin{pmatrix} \frac{(1-f)1_n}{1-N-n} \\ f \end{pmatrix},
\]

\[
\hat{\alpha} = \frac{1}{f n} Y_i^\tau, \quad V_i^\tau = f v^\tau 1_n - f^2 \frac{\sigma^2}{N} J_n, \quad v^\tau = f \sigma^2 + \sigma^2_{e\tau}, \quad \sigma^2_{e\tau} = \sigma^2 - \sigma^2_{e\tau}
\]

\[
V_{i1}^\tau = f (1-f) \left[ \sigma^2_{e\tau} \left( J_n \bigg| \begin{array}{c} 0 \\ 1_n \end{array} \right) - \frac{\sigma^2}{N} \left( J_n \bigg| \begin{array}{c} J_n \\ 1_n \end{array} \right) \otimes \left( \frac{f}{1-f} \bigg| 1 \right) \right].
\]

(see c02ed12.doc and c02ed42.doc for details). Using these expressions,

\[
\hat{P}_i = e_{i1}^T Y_i + \left( \frac{1-f}{f} \right) e_{i1}^T \left( J_n \bigg| \begin{array}{c} 0 \\ 1_n + f k^\tau P_n \end{array} \right) Y_i^\tau + \frac{1}{f} e_{i1}^T \left( J_n \bigg| \begin{array}{c} J_n \\ 1_n \end{array} \right) \otimes \left( \frac{f}{1-f} \bigg| 1 \right) Y_i^\tau
\]

(12)

where \( k^\tau = \frac{\sigma^2_{e\tau}}{v^\tau} \). When in addition, all cluster sizes are equal, (12) simplifies to

\[
\hat{P}_i = m e_{i1} Y_i + (M-m) e_{i1}^T \left[ 1_n \bar{Y} + k^\tau P_n \bar{Y} \right] + M e_{i1}^T 1_{N-n} \bar{Y}
\]

where \( \bar{Y} = \frac{\sum_{j=1}^{n} Y_{ij}}{m} \), \( \bar{Y} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n)' \), and \( \bar{Y} = \frac{\sum_{j=1}^{n} \sum_{j=1}^{m} Y_{ij}}{nm} \). (see c02ed39.doc and c02ed42.doc). The variance of (12) when \( i \leq n \) simplifies to

\[
\text{var}_{Y^\prime} (\hat{P}_i - P_i) = (1-f)^2 \left[ f (1-f) k^2 + (1-fk^\tau) \right] \left[ 1 - \frac{1}{n} \right] \sigma^2 + \frac{1-f}{n} \sigma^2_{e\tau}
\]

and when \( i > n \) reduces to \( \text{var}_{Y^\prime} (\hat{P}_i - P_i) = \sigma^2 + \frac{1}{n} \left( \sigma^2 + \frac{1-f}{f} \sigma^2_{e\tau} \right) \).

3.2. Predictors of PSU Means based on Weighted PSU Totals

Predictors of the PSU mean (13) can be developed from a collapsed set of random variables corresponding to a weighted PSU totals by using \( C_{s^\prime} = \frac{C_s}{M_s} \) in lieu of (1) to define

\[
C^\prime R^\prime = Y^\prime \text{ where } C^\prime = \left( C_1^\prime, C_2^\prime, \ldots, C_N^\prime \right). \quad \text{The elements in } Y^\prime \text{ corresponding to PSU } i
\]
are given by
\[
\sum_{i=1}^{N} U_{is} f_s \bar{Y}_{si} / \sum_{i=1}^{N} U_{is} (1 - f_s) \bar{Y}_{sit}
\].

Defining \( f = (f_1, f_2, \ldots, f_N)' \) leads to the model

(similar to (5)) that

\[
Y^* = \left[ \frac{I_N}{N} \otimes \left( f' \right) \right] \mu + E^* \tag{13}
\]

where \( E^* = C'E \). Proceeding in the same manner as in Section 3.1, the \( i^{th} \) selected PSU mean is given by \( P_i^* = g_i'R^* = g_i'Y^* \) where \( g_i' \) is given by (3). Similar developments lead to the same set of unbiased constraints given by (9) which cannot in general be satisfied. Once again, requiring second stage samples to be selected with probability proportional to size enables development of an unbiased predictor of (13).

With the assumption of probability proportion to size second stage sampling, model (13) simplifies to \( Y^* = X^* \mu + Z'B + E^* \) with variance given by

\[
\text{var}_{\psi(Y^*)} = \sigma^2_e I_N \otimes \left[ \begin{array}{cc} f & 0 \\ 0 & (1 - f) \end{array} \right] \left[ \begin{array}{c} f \\ (1 - f) \end{array} \right] - \frac{1}{N} \text{cov}(\mu, E^*)
\]

where \( \sigma^2_e = \frac{1}{N} \sum_{s=1}^{N} \sigma^2_f \).

The best linear unbiased predictor of the \( i^{th} \) selected cluster mean is given by

\[
\hat{P}_i^* = g_{ii}' Y_i^* + g_{ii}' \left[ X_i^* \hat{\alpha}^* + V_{i,ii}^{-1} \left( Y_i^* - X_i^* \hat{\alpha}^* \right) \right]
\]

where \( \hat{\alpha}^* = \frac{1}{f_n} Y_i^* \).

\[
V_i^* = f\nu' I_n - \frac{f^2}{N} J_n, \quad \sigma^2 = \sigma^2 - \sigma^2_e, \quad \nu^* = f \sigma^2 + \sigma^2_e \quad \text{and}
\]

\[
V_{i,ii} = f (1 - f) \left[ \sigma^2 (I_n + 0) \right] - \frac{\sigma^2}{N} J_n (J_n (N-n) \otimes \left[ \begin{array}{c} f \\ 1-f \end{array} \right] 1) \right)
\]

(similar to the development in c02ed12.doc and developed in c02ed42.doc). Using these expressions,

\[
\hat{P}_i^* = e_{ii}' Y_i^* + f \left[ \frac{1-f}{f} \right] e_{ii}' \left( \frac{J_n}{n} + f k' P_n \right) Y_i^* + \frac{1}{f} e_{ii}' \left( \frac{J_n (N-n)}{n} \right) Y_i^* \tag{14}
\]
where \( k^* = \frac{\sigma^2}{\nu} \). When all cluster sizes are equal, simplifies to

\[
\hat{P}_i = \frac{m}{M} \mathbf{e}_i' \overline{Y} + \frac{(M-m)}{M} \mathbf{e}_i' \left[ \mathbf{1}_n \overline{\mathbf{1}} + k^* \mathbf{1}_n \overline{Y} \right] + \mathbf{e}_i' \mathbf{1}_{N-n} \overline{Y}.
\]

The variance of (14) when \( i \leq n \) simplifies to

\[
\text{var}_{\varepsilon_i} \left( \hat{P}_i - P_i^* \right) = \left( 1 - f \right)^2 \left[ f \left( 1 - f \right) k^2 + \left( 1 - fk^* \right) \right] \left( 1 - \frac{1}{n} \right) \sigma^2 + \left( 1 - f \right) \left( 1 - \frac{1}{n} + \frac{1}{nf} \right) \sigma^2
\]

and when \( i > n \) reduces to  

\[
\text{var}_{\varepsilon_i} \left( \hat{P}_i - P_i^* \right) = \sigma^2 + \frac{1}{n} \left( \sigma^2 + \frac{1}{f} \sigma^2 \right).
\]

(see c02ed49.doc)

### 3.3. Predictors of a PSU Mean based on Sampled and Remaining SSU Averages

Predictors of a PSU mean (13) can be developed from a collapsed set of random variables corresponding to sampled and remaining SSU averages by using

\[
\mathbf{C}_{\overline{Y}} = \left( \mathbf{1}_{M_s} \otimes \mathbf{I}_N \right) \left( \begin{array}{ccc}
\frac{1}{m_s} & 0 \\
0 & \frac{1}{M_s - m_s}
\end{array} \right) \left( \begin{array}{c}
\mathbf{1}_{M_s} - m_s \\
1
\end{array} \right)
\]

in lieu of (1) to define \( \mathbf{C}^* \mathbf{R} = \mathbf{Y}^* \) where \( \mathbf{C}^* = \left( \mathbf{C}^*_{11} \; \mathbf{C}^*_{21} \; \cdots \; \mathbf{C}^*_{N1} \right) \). The elements in \( \mathbf{Y}^* \)

corresponding to PSU \( i \) are given by \( \frac{\sum_{s=1}^{N} U_{is} \overline{Y}_{isl}}{\sum_{s=1}^{N} U_{is} \overline{Y}_{isl}} \). We define a linear combination of these random variables as

\[
P_i^* = g_i^* \mathbf{Y}^* = \sum_{s=1}^{N} U_{is} \left( \mathbf{c} \overline{Y}_{isl} + \left( 1 - c \right) \overline{Y}_{isl} \right) \tag{16}
\]

where \( g_i^* = \mathbf{e}_i' \otimes \left( \mathbf{c} \mathbf{1}_s \mathbf{1}_s - c \right) \) for some constant \( c \). This random variable, although not equal to the PSU mean, has an expected value that equals the PSU mean, \( E_{g_i^*} \left( g_i^* \mathbf{Y}^* \right) = \sum_{s=1}^{N} U_{is} \mu_s \).

Notice that \( \mathbf{g}^* \mathbf{R}^* = g_i^* \mathbf{Y}^* + \sum_{s=1}^{N} U_{is} \left( \mu_s - \left( \mathbf{c} \overline{Y}_{isl} + \left( 1 - c \right) \overline{Y}_{isl} \right) \right) \). Unlike the development in sections 3.1 and 3.2, the second term in this expansion is not identically equal to zero. We ignore this term in developing a unbiased linear predictor for (16) that minimizes the expected mean squared error for any two stage sampling plan.

Using this collapsed set of random variables, model (7) reduces to
\[ Y^* = X^* \mu + Z^* B + E^* \]

where \( X^* = I_{2N} \) and \( Z^* = I_N \otimes 1_2 \), and \( E^* = C^* E \). The variance is given by

\[
\text{var}_{\hat{\mu}, \hat{\beta}}(Y^*) = I_N \otimes \begin{bmatrix} \sum_{s=1}^{N} \frac{\sigma_s^2}{Nm_s} & 0 \\ 0 & \sum_{s=1}^{N} \frac{\sigma_s^2}{N(M_s - m_s)} \end{bmatrix} + \sigma_{e^2} \left( I_N \otimes J_2 \right) - \frac{\sigma_{e^2}}{N} \left( J_N \otimes J_2 \right). 
\]

We partition these random variables into a sample and remaining vector such that

\[ K^* Y^* = \begin{pmatrix} Y_I^* \\ Y_{II}^* \end{pmatrix} \]

and similarly partition \( K^* g_i^* = \begin{pmatrix} g_{ii}^* \\ g_{iII}^* \end{pmatrix} \) such that \( g_{ii}^* = c e_i^* \) and

\[ g_{iII}^* = \left( (1-c)e_i^* - e_{II}^* \otimes (c \mid 1-c) \right). \]

Collection of the study data will result in realizing the values of \( Y_I^* \). We require the predictor of \( P_i^* \) to be a linear function of the sample data,

\[ \hat{P}_i^* = (g_{ii}^* + a') Y_I^* \]

and be unbiased such that \( E_{\hat{\mu}, \hat{\beta}}(\hat{P}_i^* - P_i^*) = 0 \). The constraint will be satisfied if \( a'I_n = 1 - c \) when \( i \leq n \), or if \( a'I_n = 1 \) when \( i > n \). The best linear unbiased predictor is given by

\[ \hat{P}^* = g_{ii}^* Y_I^* + g_{iII}^* \left[ X_{II}^* \hat{\alpha} + V_{I,II}^* V_I^{-1} (Y_I^* - X_I^* \hat{\alpha}) \right] \]

which simplifies to

\[ \hat{P}^* = c e_i^* Y_I^* + e_{II}^* \left[ (1-c) \frac{J}{n} + ck^* P_n \right] Y_I^* + e_{III}^* \left( \frac{J}{(N-n)\alpha} \right) Y_I^* \]

where \( k^* = \frac{\sigma_{e^2}}{\nu^*} \) and \( \nu^* = \sigma_{e^2} + \sigma_{e^2}^* \) and \( \sigma_{e^2} = \frac{1}{N} \sum_{s=1}^{N} \frac{\sigma_s^2}{m_s} \). The variance of (17) when \( i \leq n \) simplifies to
\[ \text{var}_{\varphi z} \left( \hat{P}_i - P^* \right) = \]
\[ = \frac{ck^* (n-1)}{n} \left[ ck^* v^* - 2(1-c)\sigma^* + (1-c) \right]^2 \left[ S^* - 2\sigma^2 + \sigma^2 + \sigma^2_{\text{eff}} \right] \]

and when \( i > n \) reduces to

\[ \text{var}_{\varphi z} \left( \hat{P}_i - P^* \right) = \sigma^2 \left( 1 - \frac{1}{N-n} \right) + \left[ v^* + 2f_n \sigma^2 + \sigma^2_{\text{eff}} - \sigma^2 \right] \left( \frac{1}{N-n} \right) + \left[ c^2 \sigma^2_{\text{eff}} + (1-c)^2 \sigma^2_{\text{eff}} \right] \]

(see c02ed51.doc) where we define \( \sigma^2_{\text{eff}} = \sum_{i=1}^{N} \frac{\sigma_i^2}{N \left( M - m_i \right)} \).

6. Application

We illustrate predictors for a selected physician’s practice in a randomized controlled trial (Ockene, Hebert et al. 1996) designed to evaluate the effectiveness of a training program for physician-delivered nutrition counseling, alone and in combination with an office support program on dietary fat intake and blood LDL-cholesterol levels in patients with hyperlipidemia (WATCH). Patients enrolled in the study had duplicate venous cholesterol measures made at baseline and at 1 year follow-up. In addition, a 24 hour dietary recall was reported at baseline and 1 year follow-up. Data from these measures, in addition to baseline questionnaire data and intervention data were used for study evaluation.

The WATCH study had a cluster randomized design, with patients in physician practices forming the basic clusters. We focus our attention here on a single dependent variable, percent of Kcal from saturated fat, which we consider as a difference constructed by taking the Baseline minus 1 year measure for each patient. We also limit discussion to a single treatment arm (III) to which fourteen physician practices were randomly assigned. A summary of the average change for different physician practices assigned to this intervention is given in Table 1. We focus on predicting response for physician practice number 71. The estimate of average change in SFAT is -1.47% Kcal. This is the estimate that results treating physician practices as fixed effects. If physician practices are considered to be random, the usual mixed model estimate of the realized random effect is -0.99% Kcal (from Proc Mixed). Using a two-stage model-based approach (Scott and Smith 1969), the best linear unbiased predictor will depend on the population size. Unfortunately, the number of eligible patients in each practice was not recorded in the study. We make two assumptions about population sizes to illustrate differences in the estimator, first assuming all physician practices see \( M = 100 \) eligible patients, and second assuming half the eligible patients in a practice are included in the study.

A mixed model was fit to these data, resulting in an estimate (REML) of the variance between MDs of 0.431. We use methods of moments estimates to develop estimates of other variance components for use in developing predictors.

| Table 1. List of Average Change in SF as Pct of Kcal by Physician for n=14 physicians in the Counseling and Support Intervention |
### Table of Average Change in Percent

<table>
<thead>
<tr>
<th># of patients</th>
<th>Percent change</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3.25</td>
<td>.</td>
</tr>
<tr>
<td>30</td>
<td>-2.31</td>
<td>35.48</td>
</tr>
<tr>
<td>11</td>
<td>-2.16</td>
<td>33.44</td>
</tr>
<tr>
<td>42</td>
<td>-2.09</td>
<td>34.18</td>
</tr>
<tr>
<td>25</td>
<td>-1.47</td>
<td>19.88</td>
</tr>
<tr>
<td>20</td>
<td>-1.38</td>
<td>21.82</td>
</tr>
<tr>
<td>25</td>
<td>-0.92</td>
<td>23.29</td>
</tr>
<tr>
<td>15</td>
<td>-0.82</td>
<td>14.59</td>
</tr>
<tr>
<td>10</td>
<td>-0.69</td>
<td>22.39</td>
</tr>
<tr>
<td>29</td>
<td>-0.46</td>
<td>18.70</td>
</tr>
<tr>
<td>15</td>
<td>0.36</td>
<td>28.73</td>
</tr>
<tr>
<td>38</td>
<td>0.37</td>
<td>27.70</td>
</tr>
<tr>
<td>16</td>
<td>0.75</td>
<td>17.53</td>
</tr>
<tr>
<td>33</td>
<td>0.91</td>
<td>20.97</td>
</tr>
</tbody>
</table>

Source: ced02p07.sas  11/16/02 by EJS

We develop predictors using variance components developed using the methods of moments corresponding to sections 3.1-3.3. The method of moment estimate of $\sigma_i^2$ is given by
$$
\hat{s}_i^2 = \frac{\sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2}{f^2(n-1)}
$$
where $Y_i = \sum_{j=1}^{m_i} Y_{ij}$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$; we use the REML of $\sigma^2$ from a mixed model; and we estimate $\sigma_{e*}^2$, $\sigma_{e}^2$, $\sigma_{a}^2$ and $\sigma_{eII}^2$ by $s_{e*}^2 = \frac{1}{n} \sum_{i=1}^{n} M_i s_i^2$, $s_{e}^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2$, $s_{a}^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2$ and $s_{eII}^2 = \frac{1}{n} \sum_{i=1}^{n} M_i s_i^2$, respectively.

Still to do: Develop example further, and check that it works. Write discussion.

Discussion

Mixed model and super-population model approaches can be directly extended to prediction in unequal size cluster sample problems. This is an attractive appeal of the approaches. However, the mixed model approaches do not account for the finite number of clusters, or differences in the number of units per cluster. While the super-population model approach can account for such differences, it requires specifying an artificial super-
population. The appeal of these approaches is the simplicity of their extension. This simplicity is achieved by distorting the problem (i.e. assuming infinite size clusters, or a super-population).
Appendix A. Evaluation of $\text{var}_{\xi|\xi'} (R^*)$

We outline the steps leading to $\text{var}_{\xi|\xi'} (R^*)$ given by (9) using the conditional expansion $\text{var}_{\xi|\xi'} (R^*) = \text{var}_{\xi} (E_{\xi|\xi'} [R^*]) + E_{\xi} (\text{var}_{\xi|\xi'} [R^*])$. Since the vector $R^*$ is constructed from sub-vectors for each unit given by $R_{st}^* = (U_s \otimes U_s^{(s)})y_{st}$, we begin by developing the expected value and variance of these sub-vectors. First,

$$E_{\xi|\xi'} (R_{st}^*) = \left( U_s \otimes \frac{1_{M_1}}{M_1} \right)y_{st} \quad \text{and thus,} \quad E_{\xi|\xi'} [R^*] = \left[ \begin{array}{c} y_1 \otimes U_1 \otimes \frac{1_{M_1}}{M_1} \\ y_2 \otimes U_2 \otimes \frac{1_{M_2}}{M_2} \\ \vdots \\ y_N \otimes U_N \otimes \frac{1_{M_N}}{M_N} \end{array} \right].$$

Now $\text{var}_{\xi} \left( \left( y_s \otimes U_s \otimes \frac{1_{M_s}}{M_s} \right) \right) = E_{\xi} \left( \left( y_s y_s' \otimes U_s U_s' \otimes \frac{1_{M_s}}{M_s} \right) \right) - \left( \left( y_s y_s' \otimes \frac{J_N}{N^2} \otimes \frac{1_{M_s}}{M_s} \right) \right)$

simplifies to $\text{var}_{\xi} \left( \left( y_s \otimes U_s \otimes \frac{1_{M_s}}{M_s} \right) \right) = \left( \left( y_s y_s' \otimes \frac{J_N}{N^2} \otimes \frac{1_{M_s}}{M_s} \right) \right)$ or

$$\text{var}_{\xi} (E_{\xi|\xi'} [R^*]) = \left( \left( y_s y_s' \otimes \frac{J_N}{N^2} \otimes \frac{1_{M_s}}{M_s} \right) \right).$$

since $E_{\xi} (U_s U_s') = \frac{I_N}{N}$ if $s = s^*$ and $E_{\xi} (U_s U_s') = \frac{1}{N(N-1)}(-I_N + J_N)$ if $s \neq s^*$. We express this as
Next we evaluate \( E \left( \text{var}_{\xi_i} \left[ R^* \right] \right) \). First, 

\[
\text{var}_{\xi_i} \left( R^*_{st} \right) = U_s U'_s \otimes \left( E_{\xi_i} \left( U^{(s)}_t U'^{(s)}_t \right) - \frac{J_{1_{M_r}}}{M^2_s} \right) y^2_{st} = \frac{1}{M_s} \left( U_s U'_s \otimes P_{M_s} \right) y^2_{st}.
\]

Also, 

\[
\text{cov}_{\xi_i} \left( R^*_{st}, R^*_{st'} \right) = U_s U'_s \otimes \left( E_{\xi_i} \left( U^{(s)}_t U'^{(s)}_t \right) - \frac{J_{1_{M_r}}}{M^2_s} \right) y_{st} y_{st'} = -\frac{1}{M_s (M_s - 1)} \left( U_s U'_s \otimes P_{M_s} \right) y_{st} y_{st'}.
\]

Using these expressions, \( \text{var}_{\xi_i} \left( R^*_{st} \right) = \frac{1}{M_s - 1} \left( \frac{M_s - 1}{M_s} y^2_{st} - \frac{y^2_{st}}{M_s} \right) \otimes \left( U_s U'_s \otimes P_{M_s} \right) \). As a result, 

\[
E \left( \text{var}_{\xi_i} \left[ R^* \right] \right) = \frac{1}{M_s - 1} \left( \frac{M_s - 1}{M_s} \right) \otimes I_{M_s} \otimes P_{M_s},
\]

and combining terms results in the expression for the variance given by (9).


structured office practice on diet and serum lipid measurements in a hyperlipidemic population.

The Worcester-area trial for counseling in hyperlipidemia (WATCH).


