Mixed models in categorical data analysis

Ruitao Zhang    Stanek Ed

1.) Non-linear model for categorical data

A finite population is defined by a listing of \( M \) days, indexed by \( t = 1, \ldots, M \) for each of \( N \) subjects, indexed by \( s = 1, \ldots, N \). We define \( y_{st} \) as the response for subject \( s \) at day \( t \). Let \( y_{st} = 1 \) represent the response ‘yes’, and \( y_{st} = 0 \) represent the response ‘no’. We parameterize response for subject \( s \) on day \( t \) as

\[
y_{st} = \pi_s + \varepsilon_{st},
\]

where, \( \pi_s = \frac{1}{M} \sum_{t=1}^{M} y_{st} \), and \( \varepsilon_{st} = y_{st} - \pi_s \) is the deviation of the response for subject \( s \) on day \( t \) from \( \pi_s \). Suppose further that we define the odds of activity as \( \pi_s/(1 - \pi_s) \), and \( \ln \left( \frac{\pi_s}{1 - \pi_s} \right) = \mu_s \). From the \( \mu_s \) defined above, we can get

\[
\pi_s = \frac{e^{\mu_s}}{1 + e^{\mu_s}}.
\]

In general, we can summarize the model as:

\[
y_{st} = f(\mu_s) + \varepsilon_{st}.
\]

Here, \( \frac{e^{\mu_s}}{1 + e^{\mu_s}} \) is non-linear function of \( \mu_s \). To estimate \( \mu_s \) in model (3), we would approximate \( f(\mu_s) \) by some methods of linearization such as Taylor’s expansion (Appendix 1).

1.) First order Taylor’s expansion:
\[ f(\mu_s) = f(\mu_n) + (\mu_s - \mu_n) \frac{\partial f(\mu_s)}{\partial \mu_{s|s=x_0}} + o(\mu_s - \mu_n) \]

\[ f(\mu_s) = f^\ast(\mu_s) + o(\mu_s - \mu_n) \]

where

\[ f^\ast(\mu_s) = f(\mu_n) + (\mu_s - \mu_n) \frac{\partial f(\mu_s)}{\partial \mu_{s|s=x_0}} \]

\[ = \frac{e^{\mu_o}}{1+e^{\mu_o}} + (\mu_s - \mu_n) \left\{ \frac{\partial e^{\mu_s}}{\partial \mu_{s|s=x_0}} \frac{1}{1+e^{\mu_o}} + \frac{\partial ((1+e^{\mu_s})^{-1})}{\partial \mu_{s|s=x_0}} e^{\mu_o} \right\} \]

here

\[ \mu_n = \ln \left( \frac{\hat{\pi}_s}{1-\hat{\pi}_s} \right) \text{ for } s=1,2,...,n, \text{ and } \mu_s = \frac{1}{n} \sum_{s=1}^{n} \ln \left( \frac{\hat{\pi}_s}{1-\hat{\pi}_s} \right) \text{ for } s=n+1,...,N. \]

\[ f^\ast(\mu_s) = \frac{e^{\mu_o}}{1+e^{\mu_o}} + (\mu_s - \mu_n) \left\{ \frac{\partial e^{\mu_s}}{\partial \mu_{s|s=x_0}} \frac{1}{1+e^{\mu_o}} + \frac{\partial ((1+e^{\mu_s})^{-1})}{\partial \mu_{s|s=x_0}} e^{\mu_o} \right\} \]

\[ = \frac{e^{\mu_o}}{1+e^{\mu_o}} + \frac{1}{(1+e^{\mu_o})^2} (\mu_s - \mu_n) \]

\[ = \frac{e^{\mu_o}}{1+e^{\mu_o}} - \frac{\mu_n}{(1+e^{\mu_o})^2} + \frac{1}{(1+e^{\mu_o})^2} \mu_s \]

Then the model becomes:

\[ y_{st} = f^\ast(\mu_s) + \varepsilon_{st} + o(\mu_s - \mu_n) \]

The above model can be approximated by the following model:

\[ y_{st}^* = f^\ast(\mu_s) + \varepsilon_{st} \]

\[ = \frac{e^{\mu_o}}{1+e^{\mu_o}} + \frac{1}{(1+e^{\mu_o})^2} (\mu_s - \mu_n) + \varepsilon_{st} \]

\[ = \frac{e^{\mu_o}}{1+e^{\mu_o}} - \frac{\mu_n}{(1+e^{\mu_o})^2} \mu_s + \frac{1}{(1+e^{\mu_o})^2} \mu_s + \varepsilon_{st} \]

\[ = k_0 + k_1 \mu_s + \varepsilon_{st} \]

where, \[ l_0 = \frac{e^{\mu_o}}{1+e^{\mu_o}} - \frac{1}{(1+e^{\mu_o})^2} \mu_s \text{ and } l_1 = \frac{1}{(1+e^{\mu_o})^2}. \]

Then the model becomes:

\[ y_{st}^* = \mu_s^* + \varepsilon_{st}, s=1,2,...,N. \]

where \[ \mu_s^* = (l_0 + l_1 \mu_s) \]
2. THE TWO-STAGE RANDOM PERMUTATION MODEL FOR A FINITE POPULATION WITH RESPONSE ERROR

The probability model we consider arises from two-stage sampling of a finite population where a simple response error model is assumed for units in the population. Sampling is incorporated into the model using indicator random variables that specify a two-stage random permutation of clusters and units in the population, and hence we refer to this model as a random permutation model. The model is similar to the super-population models discussed by Rao and Bellhouse (1978) (who also call their model a random permutation model). We first define a response error model for the population and finite population parameters. Next, we introduce sampling indicator random variables, and define target random variables. We conclude by summarizing the model along with first and second moments.

2.1. The Finite Population Response Error Model and Parameterizations

The mean and variance of the expected response for units in cluster \( s \) are defined as

\[
\mu_s = \frac{1}{M} \sum_{t=1}^{M} y_{st} \quad \text{and} \quad \sigma_s^2 = \frac{1}{M} \sum_{t=1}^{M} (y_{st} - \mu_s)^2 \quad \text{for } s = 1, \ldots, N .
\]

The parameter \( \mu_s \) is the latent value for cluster \( s \); \( \sigma_s^2 \) is the survey sampling definition of the parameter for the variance. We define the average within cluster variance as

\[
\sigma_e^2 = \frac{1}{N} \sum_{s=1}^{N} \sigma_s^2 . \quad (1.1)
\]
Similarly, we define the population mean and the between cluster variance as

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

respectively. Finally, we define

\[
\beta_s = (\mu_s - \mu) \quad \text{as the deviation of the latent value of cluster } s \quad \text{from the population mean},
\]

and \(\varepsilon_{st} = (y_{st} - \mu_s)\) as the deviation of the expected response for unit \(t\) (in cluster \(s\)) from the latent value of cluster \(s\). Using these definitions, we represent response for unit \(t\) in cluster \(s\) as

\[
y_{st} = \mu + \beta_s + \varepsilon_{st}. \tag{1.2}
\]

Model (2.4) is called a derived model (Hinkelmann and Kempthorne 1994). We represent model (2.4) simultaneously for all units in the finite population using matrix notation. Defining \(y = \left( y_1' \quad y_2' \quad \cdots \quad y_N' \right)'\) where

\[
y_s = \left( y_{s1} \quad y_{s2} \quad \cdots \quad y_{sM} \right)',
\]

model (2.4) can be summarized as

\[
y = X\mu + Z\beta + \varepsilon \tag{1.3}
\]

where

\[
X = I_N \otimes 1_M, \tag{1.4}
\]

\[
Z = I_N \otimes 1_M, \tag{1.5}
\]

\[
\beta' = (\beta_1 \quad \beta_2 \quad \cdots \quad \beta_N), \tag{1.6}
\]

\(1_a\) 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 (2.5) are random variables.

2.2. Random Variables and the Two-Stage Random Permutation Model
We define the two-stage random permutation model by an ordered list of \( NM \) random variables, where both clusters and units in clusters are independently permuted. Assuming that each realization of the two-stage 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 label the positions in the permutation of units in a cluster by \( j = 1, \ldots, M \). Since any unit in any cluster may occupy position \( ij \), we represent the expected response (over response error) for the \( j^{th} \) SSU in the \( i^{th} \) PSU as the random variable \( Y_{ij} \). For ease of exposition, we refer to the cluster that will occupy position \( i \) in the permutation of clusters as the \( i^{th} \) primary sampling unit (PSU), and to the unit that will occupy position \( j \) in the permutation of units within a cluster as the \( j^{th} \) secondary sampling unit (SSU). 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.

To relate \( y_{st} \) to \( Y_{ij} \), we use sampling indicator random variables, \( U_{is} \), which take on a value of one when the \( i^{th} \) PSU is cluster \( s \), and a value of zero otherwise; and indicator random variables \( U_{jt}^{(s)} \) that take on a value of one when the \( j^{th} \) SSU in cluster \( s \) is unit \( t \), and zero otherwise. As a consequence, the random variable corresponding to the \( i^{th} \) PSU and the \( j^{th} \) SSU within the \( i^{th} \) PSU in a permutation is given by

\[
Y_{ij} = \sum_{s=1}^{N} \sum_{t=1}^{M} U_{is} U_{jt}^{(s)} y_{st} \cdot \tag{1.7}
\]
Letting $U^{(s)} = (U_1^{(s)} \quad U_2^{(s)} \quad \cdots \quad U_M^{(s)})$ denote the $M \times M$ matrix with columns $U_t^{(s)} = (U_{1t}^{(s)} \quad U_{2t}^{(s)} \quad \cdots \quad U_{Mt}^{(s)})'$, $U = (U_1 \quad U_2 \quad \cdots \quad U_N)$ denote the $N \times N$ matrix with columns $U_s = (U_{1s} \quad U_{2s} \quad \cdots \quad U_{Ns})'$, and $Y = \left( (Y_{ij}) \right) = \left( Y_1' \quad Y_2' \quad \cdots \quad Y_N' \right)'$ denote an $NM \times 1$ column vector with $Y_i = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{iM})'$, a vector of random variables defining the two stage permutation of the population is

$$Y = \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^N U^{(s)} \right) y \quad (1.8)$$

where $\bigoplus_{s=1}^N A_s$ denotes a block diagonal matrix with blocks $A_s$ (Harville 1997).

We use the parameterization in (2.4) and relationship (2.10) to express the random permutation model as a mixed model. Using the elementary properties of the indicator random variables, it follows that $U_1^N = I_N^N$, $U^{(s)} I_M = 1_M$, $(U \otimes I_M) \left( \bigoplus_{s=1}^N U^{(s)} \right) X = X$, and $(U \otimes I_M) \left( \bigoplus_{s=1}^N U^{(s)} \right) Z = U \otimes I_M = ZU$. The two stage permutation of $X$ is non-stochastic due to the simple structure of $X$ in (2.6), while the two stage permutation of $Z$ can be factored into the product of a non-stochastic matrix times $U$ due to the simple structure of $Z$ in (2.7). Multiplying $U$ by $\beta$, we define

$$B = U \beta \quad (1.9)$$

where $B = (B_1 \quad B_2 \quad \cdots \quad B_N)'$. The terms $B_i = \sum_{s=1}^N U_{is} \beta_s$ for $i = 1, \ldots, N$ are random effects, and represent the deviation of the latent value for PSU $i$ from the population mean. Combining these expressions, and defining $E = \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^N U^{(s)} \right) \epsilon$, we obtain the random permutation mixed model

$$Y^* = X \mu + ZB + E \quad (1.10)$$

Model (2.14) differs from the usual mixed model since it represents all elementary units in the population, as opposed to solely sampled units. The term $E$ represents the deviations of response for the SSUs from the PSU latent values. The average of the first $M$ random variables in $Y$ is the mean for the first PSU. A realization of $Y$ corresponds to a particular two-stage permutation of the population. The vector $Y$ is exchangeable.
2.3. Target Random Variables

We assume that there is interest in a linear combination of the random variables \( Y \) of the form

\[
T = g'Y
\]  

(1.11)

where \( g' = (g_1', g_2', \ldots, g_N') \) with \( g_i = (g_{i1}, \ldots, g_{iM})' \), \( i = 1, \ldots, N \), is a vector of known constants. Linear combinations of PSU means may be defined by taking

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

(1.12)

where \( b = (b_1, \ldots, b_N)' \) is a vector of known constants. Of principal interest is the linear combination that defines the latent value of PSU \( i \), i.e.,

\[
g' = e_i' \otimes \frac{1_M}{M}, \]  

(1.13)

with \( e_i \) denoting an \( N \times 1 \) vector with a value of one in position \( i \), and zero elsewhere.

From (2.14) and (2.17), it follows that

\[
T = \sum_{s=1}^{N} U_{is} \mu_s. \]  

(1.14)

In the context of the super-population model described by Scott and Smith (1969), extended to include response error by Bolfarine and Zacks (1992), the mean of PSU \( i \) is defined as

\[
T_A = g'Y^* \]  

(1.15)

where \( T_A = T + g'W^* \) and \( g' \) is defined by (2.17). The random variables (2.15) and (2.19) have different interpretations, with the difference being the addition of a response error term. A strength of the finite population sampling based approach is the ability to distinguish such subtleties in target random variables.
2.4. First and Second Moments of the Random Permutation Mixed Model

Under the two-stage random permutation model (2.14), 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 units in a cluster, and the subscript $\xi_3$ to denote expectation with respect to response error, 

$$E_{\xi_1 \xi_2 \xi_3} (Y^*) = X\mu \quad (1.16)$$

and

$$\text{var}_{\xi_1 \xi_2 \xi_3} (Y^*) = \left( \sigma_r^2 + \sigma_e^2 \right) I_{NM} + \sigma_r^2 \left( I_N \otimes J_M \right) - \frac{\sigma_e^2}{N} J_{NM} \quad (1.17)$$

where

$$\sigma_e^2 = \sigma^2 - \frac{\sigma_e^2}{M} \quad (1.18)$$

and $J_e$ denotes an $(a \times a)$ matrix with all elements equal to one (see Appendix A).

3. PREDICTING THE MEAN AND LATENT VALUE OF A PSU BASED ON A TWO-STAGE SAMPLE

We assume that $m$ units in each of $n$ clusters are selected by two-stage sampling and that a single measure is obtained on each of the $nm$ SSUs in the sample. The process of sampling gives rise directly to the stochastic representation (2.14) of the population. The random variables in the sample will be realized. Although our principal interest is in predicting the latent value of a realized PSU defined by (2.15) and (2.17),
we develop unbiased predictors for more general linear combinations of random variables. The predictors are linear functions of the random variables in the sample that 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, and Royall 2000). We first partition the elements of $Y^*$ into a sampled and a remaining portion \[ \begin{pmatrix} Y_1^* \\ Y_2^* \end{pmatrix} = \begin{pmatrix} K, Y^* \\ K_Y Y^* \end{pmatrix}, \]
pre-multiplying $Y^*$ by

\[ K = (K_1' | K_2')' \]
where $K_1 = \begin{pmatrix} I_n & 0 \\ 0 & I_m \end{pmatrix}$, and

\[ K_2 = \begin{pmatrix} 0 & I_{M-m} \\ I_{N-n} & 0 \end{pmatrix}. \]

We partition $X$ and $Y$ in a similar manner,

such that \[ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} K_X, X \\ K_Y Y \end{pmatrix}, \]

\[ \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} K_{1, Y} Y \\ K_{2, Y} Y \end{pmatrix}, \]

\[ \text{var}_{Y^*} \left( \begin{pmatrix} Y_1^* \\ Y_2^* \end{pmatrix} \right) = \text{var}_{Y^*} \left( \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \right) + \sigma^2 I_{NM-nm} \]
where \[ \text{var}_{Y^*} \left( \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \right) = \begin{pmatrix} V_1 & V_{1,2} \\ V_{1,2}' & V_2 \end{pmatrix}, \]

\[ V_1 = \sigma^2 I_{nm} + \sigma^2 \left( I_n \otimes J_{m} \right) - \frac{\sigma^2}{N} J_{nm}, \]

\[ V_{1,2} = \sigma^2 \left( I_n \otimes J_{m} \right) \otimes I_{nm} - \frac{\sigma^2}{N} J_{nm} \]

and \[ V_2 = \sigma^2 I_{NM-nm} + \sigma^2 \left( I_n \otimes J_{M-m} \right) \otimes I_{N-n} - \frac{\sigma^2}{N} J_{NM-n}. \]

Finally, we partition $g'$ resulting in \[ \left( g_1' \mid g_2' \right) = \left( g'K_1' \mid g'K_2' \right) \]
and partition $g''_2$ corresponding to $K_2$ as

\[ g''_2 = \left( g_{1,2}' \mid g_{2,2}' \right). \]
We assume that the subset \( Y^* \) of the elements of \( Y \) will be realized, call such a subset "the sample", and express the target random variable \( T \) (or \( T' \)) as the sum of two parts. One part depends on the potentially realized sample; the other depends on the remaining random variables. We require the predictor of (2.15) to be a linear function of the sampled random variables, i.e., that \( \hat{T} = (g'_I + a^I) Y^*_I \), to be unbiased, resulting in the constraint that \( a'^I X_I - g'^I X_{II} = 0 \), and to minimize the expected value of the mean squared error given by \( \text{var}_{\tilde{\sigma}_2} (\hat{T} - T) \). Notice that the unbiased constraint only requires
\[
E_{\tilde{\sigma}_2} (\hat{T}) = E_{\tilde{\sigma}_2} (T) \text{ (see also Robinson 1991)}.
\]
Since cluster sizes and sample sizes within clusters are equal, and a single observation is included in the model for each SSU, the matrices \( X_I \) and \( X_{II} \) are non-stochastic, avoiding the problem noted by Pfeffermann (1984) in Porter’s (1973) development of predictors. Letting
\[
\begin{align*}
\hat{T} - T &= \left( (a' + g'_I) \right) \begin{pmatrix} Y^*_I \\ Y_I \\ Y_{II} \end{pmatrix} - \begin{pmatrix} -g'_I \\ -g'^I \end{pmatrix} \begin{pmatrix} Y^*_I \\ Y_I \\ Y_{II} \end{pmatrix}
\end{align*}
\]
and noting that
\[
\text{var}_{\tilde{\sigma}_2} \begin{pmatrix} Y^*_I \\ Y_I \\ Y_{II} \end{pmatrix} = \begin{pmatrix} V^*_I & V_I & V_{I,II} \\ V_I & V_I & V_{I,II} \\ V_{II,I} & V_{II,J} & V_{II} \end{pmatrix}
\]
where \( V^*_I = V_I + \sigma^2 I_{nm} \), minimization of
\[
\text{var}_{\tilde{\sigma}_2} (\hat{T} - T) \text{ subject to the unbiased constraint using Lagrangian multipliers leads to}
\]
\[
\hat{T} = g'_I \begin{pmatrix} X_I \hat{\alpha} + V_I V_I^{-1} (Y^*_I - X_I \hat{\alpha}) \end{pmatrix} + g'^I \begin{pmatrix} X_{II} \hat{\alpha} + V_{I,II} V_I^{-1} (Y^*_I - X_I \hat{\alpha}) \end{pmatrix}
\] (2.1)
where \( \hat{\alpha} = \left( X_i' V_i^{-1} X_i \right)^{-1} X_i' V_i^{-1} Y_i^* \). Simplifying terms, \( \hat{\alpha} = Y_{nm}^* Y_i^* = \mathbf{Y}^* \),

\[
Y_i^* - X_i \hat{\alpha} = P_{nm} Y_i^*, \quad V_i' V_i^{-1} P_{nm} = \left[ \rho_i I_{nm} + (1 - \rho_i) k^* \left( I_n \otimes \frac{J_m}{m} \right) \right] P_{nm}
\]

and

\[
V_{i,li} ' V_i^{-1} P_{nm} = k^* \begin{bmatrix}
I_n \otimes \frac{J}{(M-m) \times m} \\
0 \\
0
\end{bmatrix}
P_{nm}
\]

where \( \rho_i = \frac{\sigma^2}{\sigma^2 + \sigma^2_r} \), \( k^* = \frac{m \sigma^2}{m \sigma^2 + (\sigma^2 + \sigma^2_r)} \), and

\( P_a = I_a - \frac{J_a}{a} \). Using these expressions, the predictor simplifies to

\[
\hat{T} = g_l' \left[ \frac{J_{nm}}{nm} + \left( \rho_i I_{nm} + (1 - \rho_i) k^* \left( I_n \otimes \frac{J_m}{m} \right) \right) P_{nm} \right] Y_i^*
\]

\[
+ g_{l,li}' \left[ \frac{J_{nm}}{nm} + k^* \begin{bmatrix}
I_n \otimes \frac{J}{(M-m) \times nm} \\
0 \\
0
\end{bmatrix}
P_{nm} \right] Y_i^*
\]

which further simplifies to

\[
\hat{T} = g_l' \left[ I_{nm} \bar{Y}_l^* + (1 - \rho_i) k^* \left( P_n Y_i^* \otimes I_m \right) + \rho_i P_{nm} Y_i^* \right]
\]

\[
+ g_{l,li}' \left[ I_{nm} \bar{Y}_l^* + k^* \left( P_n \bar{Y}_l^* \otimes I_{M-m} \right) \right] + g_{2,li}' \left[ I_{N-n(M-m)} \bar{Y}_l^* \right]
\]

where \( \bar{Y}_l^* = \left( I_n \otimes \frac{I_m}{m} \right) Y_l^* \) is a vector of PSU sample means. The first term in equation (3.2) is the predictor of the linear combination of random variables associated with the realized SSUs for the sample PSUs. The second term is the predictor of the linear combination of SSUs that were not realized for the realized PSUs, while the third term is the predictor of the linear combination of the SSUs for PSUs not in the sample. Defining \( g' \) as in (2.17), the predictor in (3.2) simplifies to
\[
\hat{T} = \left( \frac{m}{M} \right) e'_i \left( I_n \bar{Y}_i + \left[ \rho_i + (1 - \rho_i) k^* \right] P_n \bar{Y}_i^* \right) + \left( \frac{M - m}{M} \right) e'_i \left( I_n \bar{Y}_i + k^* P_n \bar{Y}_i^* \right) + (e'_{iN} 1_{N-n}) \bar{Y}_i^*,
\]

(2.3)

where \( e'_i = \begin{pmatrix} e'_{iN} \\ e'_{iI} \end{pmatrix} \). The first two terms in (3.3) correspond to the predictor of the linear combination of the random variables associated with a sample PSU (where \( i \leq n \)); the third term is the predictor of the linear combination of the random variables associated with a PSU that is not in the sample (\( i > n \)). The term \( \rho_i + (1 - \rho_i) k^* \) accounts for response error, and shrinks the realized values of the random variables associated with the sample SSUs towards the sample mean when \( i \leq n \). The term \( k^* \) accounts for both response error and between SSU variation, and shrinks the predictors of the random variables associated with the SSUs not included in the sample towards the sample mean. A similar result holds for more general linear combinations of random variables associated with PSU latent values where \( e'_i = \begin{pmatrix} e'_{iN} \\ e'_{iI} \end{pmatrix} \) is replaced by \( b' = \begin{pmatrix} b'_I \\ b'_{iN} \end{pmatrix} \) in (3.3).

We compare the predictor of (2.15) given by (3.2) to the predictor of (2.19) given by

\[
\hat{T}_A = g'_i Y^*_i + g'_{iI} \left[ I_{n(M-m)} \bar{Y}_i^* + k^* \left( P_n \bar{Y}_i^* \otimes I_{M-m} \right) \right] + g'_{2I} \left[ I_{(N-n)M} \bar{Y}_i^* \right].
\]

(2.4)

The predictors differ only in the first term. Since the target random variable defined by (2.15) includes response error, the values associated with the realized SSUs are used directly in the predictor. When \( g' \) is defined by (2.17), the predictor (3.4) simplifies to
This predictor is nearly identical to the predictor given by Bolfastine and Zacks (1992), apart from a difference in PSU variance parameter definitions.

When there is no response error, $\sigma_i^2 = 0$, and the target random variables given by (2.15) and (2.19) are the same. In this case, the predictors given by (3.3) (or (3.5)) simplify to

$$\hat{T} = \left( \frac{m}{M} \right) e'_{i|} \bar{Y}_i + \left( \frac{M - m}{M} \right) e'_{i|} \left( 1_n \bar{Y} + k P_n \bar{Y}_i \right) + \left[ e'_{i|} 1_{N-n} \right] \bar{Y} \quad (2.6)$$

since $e'_{i|} \left( 1_n \bar{Y} + P_n \bar{Y}_i \right) = e'_{i|} \bar{Y}_i$, where $\bar{Y}_i = \frac{\sum_{j=1}^{m} Y_{ij}}{m}, \bar{Y}_i = (\bar{y}_1 \bar{y}_2 \cdots \bar{y}_n)', \bar{Y} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}}{nm}$,

and $k = \frac{m \sigma^2 + \sigma_e^2}{m \sigma^2 + \sigma_e^2}$. 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 defined by (2.22) as opposed to $\delta^2$. Although the predictors given by (1.3) and (3.6) are nearly identical, the model assumptions differ. The predictor (3.6) is developed assuming a response error model for each unit, and a two-stage random permutation of clusters and units in the finite population. These assumptions give rise to different first and second moments than the super-population model assumptions that underlie (1.3). The assumptions required for (3.6) represent the sampling design and may be more readily accepted.

The expected MSE of the best linear unbiased predictors can be developed in a straightforward manner (see Appendix B). When predicting a target random variable
defined by (2.15) with \( g' \) defined by (2.17), the expression for the expected MSE simplifies when \( i \leq n \) to

\[
MSE(\hat{T}) = (1 - f \rho) \left[ \frac{\sigma^2}{nm \rho} + \left( \frac{n - 1}{n} \right) \left( 1 - k' \right) \sigma^2 \right]
\]

(2.7)

and when \( i > n \) to

\[
MSE(\hat{T}) = \sigma^2 \left( 1 - \frac{1}{N - n} \right) + \frac{\sigma^2}{n} + \frac{m \sigma^2 + \sigma^2 + \sigma^2}{nm}, \text{ where } f = \frac{m}{M}.
\]

In practice, variance parameters will be unknown and need to be estimated. Method of moment estimates can be obtained by equating the expected mean squares to the observed mean squares between clusters (MSB) and the observed mean squared error (MSE) from an ANOVA table, assuming \( \sigma^2 \) is known. We replace negative variance estimates by zero. The estimate of the within cluster variance is given by

\[
\hat{\sigma}_e^2 = \max \left( 0, MSE - \sigma^2 \right).
\]

The between cluster variance is estimated by

\[
\hat{\sigma}^2 = \max \left( 0, \frac{1}{m} \left[ MSB - MSE + f \hat{\sigma}_e^2 \right] \right) \text{ and we estimate } \sigma^* \text{ by}
\]

\[
\hat{\sigma}^* = \max \left( 0, \frac{1}{m} \left[ MSB - MSE \right] \right).
\]

The MSE of predictors that use estimated variance components will underestimate the true MSE since they will not account for variability in the variance estimates in the predictor. A small scale simulation study illustrated that the underestimation may be substantial, and presents a complex pattern motivating the need for additional work in this area.

4. COMPARISON OF PREDICTORS

The predictors developed in Section 3 can be compared directly with predictors developed under mixed model or super-population model assumptions. Using common
notation and setting \( k^*_i = \frac{m \sigma^2 + \sigma^2_v}{m \sigma^2 + \sigma^2_v + \sigma^2_r} \), when \( i \leq n \), the predictors of the latent value of PSU \( i \) defined by (2.15) and (2.17) under different models are summarized in Table 1.

Table 1. Predictors of the Latent Value of PSU \( i \) when \( i \leq n \) in Two-stage Cluster Sampling

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed Model</td>
<td>( \hat{\rho} = \mu + k_i (\bar{Y}_i^* - \hat{\mu}) )</td>
</tr>
<tr>
<td>Scott&amp;Smith</td>
<td>( \hat{P}_i = f \bar{Y}_i^* + (1-f) (\hat{\mu} + k_i (\bar{Y}_i^* - \hat{\mu})) )</td>
</tr>
<tr>
<td>Random. Perm.</td>
<td>( \hat{T}_i = f \bar{Y}_i^* + (1-f) (\bar{Y}_i^* + k_i (\bar{Y}_i^* - \bar{Y}^*)) )</td>
</tr>
<tr>
<td>RP + Resp. Err.</td>
<td>( \hat{T}_i = f \left( \bar{Y} + k_i (\bar{Y}_i^* - \bar{Y}^<em>) \right) + (1-f) (\bar{Y} + k_i (\bar{Y}_i^</em> - \bar{Y}^*)) )</td>
</tr>
</tbody>
</table>

Each predictor can be interpreted as the weighted sum of a term predicting the latent values of the SSUs in the sample, and a term predicting the latent values for the remaining SSUs. The mixed model predictor places all the weight on the second term, and is the limit of the predictor proposed by Scott and Smith (1969) under a super-population model when the fraction of SSUs observed on a realized PSU is so small that it can basically be neglected. In the balanced two-stage sampling setting considered here, the predictor developed by Scott and Smith (1969) is nearly identical to the predictor based on the random permutation model. The difference between the predictors results from different definitions of variance components and shrinkage constants. The two-stage random permutation model permutes PSUs, and hence has a single SSU component of variance representing the average of the SSU within cluster variances (2.3) (as opposed to the cluster specific components, \( \sigma^2_i \)). This, along with the use of a between PSU component of variance defined by (2.22) as opposed to \( \delta^2 \), results in the different predictors. The context described by Scott and Smith (1969) appears to match the
context for two-stage sampling, but their model assumptions for the variance components
do not correspond to the variance components that arise from permuting clusters and
units in a finite population.

When response error is present, the mixed model and super-population models
predictors remain the same, but the random permutation model predictor changes. Under
the random permutation model with response error, the linear combination of the latent
values of the realized SSUs for a selected PSU is predicted using a shrinkage constant,
k^*_r, determined by the response error variance.

We simplify the expressions for the predictors in Table 1 under the additional
assumptions that the within cluster variance is identical for all clusters (and equal to \( \sigma^2_w \)),
and that the response error variance is identical for all units (and equal to \( \sigma^2_r \)). With
these assumptions, \( \sigma^2_i = \sigma^2_w + \sigma^2_r \). Assuming in addition that \( \delta^2 = \sigma^2 \), then

\[
k_i = k^*_i = \frac{m \sigma^2}{m \sigma^2 + \sigma^2_w + \sigma^2_r},
\]

and each predictor in Table 1 can be expressed as

\[
\hat{T} = \overline{Y}^* + c\left(\overline{Y}_i - \overline{Y}^*\right),
\]

where the shrinkage constants \( c \) are given in Table 2. When
expressed in this form, the difference in the predictors results from the different shrinkage
constants. Since values of \( k_i, k^*, \), and \( \rho_i \) are all between zero and one, each of the
predictors will be closer to the overall sample mean than the simple mean of the realized
SSUs for a realized PSU.
Table 2. Values of $c$ for Predictors $\hat{T} = \bar{Y} + c(\bar{Y}_i - \bar{Y})$ of the Latent Value of PSU $i$ when $i \leq n$ in Two-stage Cluster Sampling with Homogeneous Unit and Response Error Variances.

**Model**

- **Mixed Model**
  
  \[ c_{MM} = \frac{k_i}{k} \]

- **Scott & Smith**

  \[ c_{SS} = f + (1 - f)k_i \]

- **Random Permutation**

  \[ c_{RP} = f + (1 - f)k \]

- **Random Permutation with Response Error**

  \[ c_{RPR} = f\rho_i + (1 - f\rho_i)k^* \]

Under the random permutation model, the expected MSE of each of the predictors can be expressed as a quadratic function of $c$ given by

\[
MSE(\hat{T}) = (1 - f\rho_i) \left[ \frac{\sigma^2}{nm\rho_i} + \left(\frac{n-1}{n}\right) (1 - k^*) \sigma^2 \right] \\
+ \left(\frac{n-1}{n}\right) \frac{\sigma^2}{k^*} \left( c - \left[ f\rho_i + (1 - f\rho_i)k^* \right] \right)^2.
\]  

(3.1)

For the random permutation model with response error, the expected MSE simplifies to (3.7).

We can use the expression for the expected MSE to compare the expected MSE of different predictors under the two stage random permutation model. The difference in expected MSE between the mixed model and the random permutation model predictor is given by

\[
MSE(\hat{T}_{MM}) - MSE(\hat{T}_{RPR}) = \left(\frac{n-1}{n}\right) \frac{\sigma^2}{k^*} (c_{MM} - c_{RPR})^2,
\]

while the difference in expected MSE between Scott and Smith’s predictor and the random permutation model predictor is given by

\[
MSE(\hat{T}_{SS}) - MSE(\hat{T}_{RPR}) = \left(\frac{n-1}{n}\right) \frac{\sigma^2}{k^*} (c_{SS} - c_{RPR})^2.
\]

Since each
expected MSE is evaluated under the random permutation model assumptions where $\hat{T}_{RPR}$ is optimal, the differences are always positive. When evaluated under the random permutation model, neither the mixed model predictor nor Scott and Smith’s predictor will have uniformly smaller expected MSE.

Figure 1 illustrates the percent increase in the expected MSE that would occur when using the mixed model or Scott and Smith’s predictor under a two stage random permutation model. In certain settings, the expected MSE for the mixed model and Scott and Smith’s predictor will exceed that of the random permutation model predictor by more than 100%. The plots in Figure 1 are organized by columns with increasing sampling fractions ($f$) for a PSU. Rows correspond to increasing cluster intra-class correlations, $\rho_c = \frac{\sigma^2}{\sigma^2 + \sigma_e^2}$. The abscissa for the individual plots is the unit intra-class correlation, $\rho_i$.

Several patterns emerge from Figure 1. First, Scott and Smith’s predictor has high expected MSE for small unit intra-class correlations ($\rho_i$), especially when the intra-class correlation is modest ($\rho_i < 0.5$). Mixed model predictors have high expected MSE when the SSU sampling fraction is large, the unit intra-class correlation is large, and the cluster intra-class correlation is small. When unit intra-class correlations are large ($\rho_i > 0.75$), Scott and Smith’s predictor has smaller expected MSE compared with the mixed model predictor. When the SSU sampling fraction is small ($f < 0.1$), the difference in expected MSE between the mixed model and the random permutation model predictor is small (less than 5%).
5. EXAMPLE

We consider a simple example from a study of seasonal variation in serum cholesterol which we refer to as the Seasons study (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 20-70 year old members of the Fallon Health Maintenance Organization (HMO) based on 5000 patient contacts (Ockene et al. 2003) to quantify seasonal patterns in cholesterol. Triplicate 24-hour diet recalls were collected prior to fasting lipid measures to control for the impact of diet on cholesterol. A total of 414 subjects had three 24-hour dietary recalls collected in a 6 week period prior to the first fasting cholesterol measure. The days were selected using a stratified random sample of days in the eligible period, with two weekday and one weekend day selected. We focus on estimating saturated fat intake for a subject (cluster) corresponding to the first measure of serum cholesterol based on the three 24 hour diet recalls in the 6 week period preceding the cholesterol measure.

Neither the study subjects (which we consider here as clusters), nor the 24-hour recall days were selected via simple random sampling. Assuming the 20-70 year old Fallon HMO membership is 25,000, we estimate the volunteer ‘participatory’ population to be of size \( N = 2070 \). A stochastic model is needed to make predictions. We assume a two stage random permutation model to predict average latent saturated fat intake for sampled subjects over the 6 week (M=42 day) period. The model is simple and plausible, but ignores the stratification of weekdays and weekends and assumes that the participating subjects are selected via simple random sampling.
The average saturated fat intake for the $n = 414$ study subjects was 24.8 g/d, with individual subject averages ranging from 2.7 g/d to 92 g/d. The 95th percentile average saturated fat intake was equal to 50 g/d. A one way ANOVA model resulted in $MSB = 484 \cdot (g/d)^2$, and $MSE = 165 \cdot (g/d)^2$. Unfortunately, the response error variance for saturated fat intake based on a 24-hour recall is not known. Reliability, i.e. $\rho_t$, is estimated to range from about 0.4 to 0.9 ((Willett 1990) based on dual 24-hour interviews on the same day. Using these estimates, we assume that the response error variance is $\sigma_r^2 = 25 \cdot (g/d)^2$ or $\sigma_r^2 = 100 \cdot (g/d)^2$ when evaluating the predictors.

For illustration, we predict the latent saturated fat intake over the 42 day period for the realized subject whose average 3-day saturated fat intake was 50 g/d. Using method of moments estimates, when $\sigma_r^2 = 25 \cdot (g/d)^2$, variance components are estimated as $\hat{\sigma}^2 = 110 \cdot (g/d)^2$ and $\hat{\sigma}_e^2 = 140 \cdot (g/d)^2$. Using the variance components to estimate $c$ in Table 2, $\hat{T}_{RPR} = 41.9 \cdot (g/d)$. Predictors based on a mixed model ($\hat{T}_{MM} = 41.5 \cdot (g/d)$) or Scott and Smith’s model ($\hat{T}_{SS} = 42.2 \cdot (g/d)$) are similar. The expected MSEs of the predictors are also similar, since $f = 0.071$ and estimates of $\rho_t$ and $\rho_r$ are given by 0.44 and 0.85, respectively (see Figure 1).

Somewhat larger differences occur when we predict the latent saturated fat intake over a 7 day period for the realized subject whose average 3-day saturated fat intake was 50 g/d assuming $\sigma_r^2 = 100 \cdot (g/d)^2$. Now, $\hat{\sigma}^2 = 116 \cdot (g/d)^2$ and $\hat{\sigma}_e^2 = 65 \cdot (g/d)^2$, so that $\hat{T}_{RPR} = 42.9 \cdot (g/d)$. The mixed model predictor is $\hat{T}_{MM} = 41.9 \cdot (g/d)$ and Scott and Smith’s predictor is $\hat{T}_{SS} = 45.3 \cdot (g/d)$. The expected MSEs of the predictors are once
again similar, since \( f = 0.428 \) and estimates of \( \rho_s \) and \( \rho_i \) are given by 0.64 and 0.39, respectively (see Figure 1).

6. DISCUSSION

We develop a predictor of latent values of clusters in a finite population based on a simple probability model that combines a response error model for each unit, and a random permutation model for the two-stage finite population sampling. The model is based on the design, explicitly accounting for the labeled finite population of clusters and units.

Several other stochastic models have been proposed in this setting. The usual mixed model sidesteps defining a finite population, or clusters and units, while accounting for a hierarchy of random variables via an assumed mean and variance structure for the sample. Model based survey methods attempt to capture the concept of a finite population by adding random variables that represent the non-sampled clusters and units. The combination of these random variables and the sample random variables constitute a super-population. A finite population is defined as the realization of the super-population random variables.

Each of the models can be characterized by the first and second moments of a set of random variables (\( nm \) random variables in the mixed model, and \( NM \) random variables in the random permutation and super-population models). Predictors are developed in a similar manner in all three models. However, the latent value of a PSU is not readily specified in the super-population model when response error is present, and
the mixed model does not account for finite population sampling of units. In contrast, the random permutation model with response error can be used to specify the latent value of a PSU while accounting for finite population sampling.

Appealing to the advantages of the random permutation model, we compare the predictors of the latent value for PSU \( i \) (when \( i \leq n \)) in Table 2 of Section 4. If we view the mixed model and super-population model as earlier attempts to mimic the two stage sampling, it is appropriate to evaluate predictors from these models under the random permutation model assumptions. Statistical inference boils down to predicting linear combinations of latent values of units, either observed with error, or unobserved. The BLUP in mixed models (1.2) is the limit of the predictor proposed by Scott and Smith (1969) under a super-population model when the fraction of SSUs observed in a realized PSU is so small that it can basically be neglected. The traditional BLUP places all the weight on predicting the random variables associated with the unobserved SSUs. Ignoring the proportion of observed SSUs has the disadvantage of increased model sensitivity due to reliance solely on model based predictors. For example, when there is no response error, if a high fraction of SSUs, say 90%, are observed for a selected PSU, a large portion of the units comprising the PSU are known, and need not be predicted. Mixed model approaches act as if all SSUs must be predicted.

In the balanced two-stage sampling setting considered here, the predictor developed by Scott and Smith (1969) is nearly identical to the predictor (3.6) based on the finite population sampling model. The context described by Scott and Smith appears to match the context for the two-stage sampling, but the super-population model assumptions for the variance components do not correspond to the two-stage sampling
variance components. It is difficult to conceive of a data generation scheme with clusters and units that would lead to the super-population model assumptions. The super-population model is in this sense artificial. When response error is present, different target random variables can be specified corresponding to the latent value of a PSU (2.18) that are not captured in the super-population model framework.

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 co-variables 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 the number of measures per SSU is part of the sampling design.

Extensions to settings where there are different cluster sizes, and possibly different sampling fractions for different clusters are more complicated. For such extensions when there is no response error, as shown in (Stanek and Singer 2002), basic random variables are required that identify both units and positions, i.e. $Y_{uij} = U_{ui} U_j(y_{ij}) y_{ui}$. Such expanded random variable notation enables the nesting of subjects in clusters to be preserved in a random permutation model. Difficulties arise with the expanded random variables due to singularities that have led to representing data for each PSU in the sample and remainder as totals, or averages. Finally, extensions to settings involving covariables either on PSUs and/or SSUs are of obvious interest. Such extensions have
been developed in a related context with domains, but not extended to a two-stage sampling setting (Lencina 2002; Li 2003).

A final caution is warranted for the interpretation of predictors of random effects. An appealing interpretation of the predictor of a realized random effect is that it predicts the latent value for the realized PSU. Thus, if the realized value of \( U_{is} \) is equal to one, the target parameter given by (2.18) is \( \mu_i \). This interpretation, while seemingly self-evident, may not apply in a two-stage sampling setting. A hint of some problem is apparent in the fact that clusters are not identifiable when representing the two-stage random permutation model as in (2.10). Dividing the random variables into a sample and remaining portion (as in Section 3) will identify clusters that correspond to realized PSUs, since the values of \( U_{is} \) will be realized (for \( i \leq n \)) in the sample. However, among the remaining random variables, the realized PSU cannot be identified, although the position of the PSU is known. When PSUs are of different size, some extra care is needed (beyond the usual representation as in (2.10)) to verify that the predicted SSUs actually correspond to SSUs for the realized PSU. Aspects of the distinction between a cluster, and a PSU are discussed in the context of simple random sampling elsewhere (Lencina 2002; Stanek, Singer and Lencina 2003). 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, i.e. $\text{var}_{\xi_1 \xi_2}^1 (Y^*) = E_{\xi_1 \xi_2} \left[ \text{var}_{\xi_1 \xi_2} (Y^*) + \text{var}_{\xi_1 \xi_2} (Y^*) \right]$. Given $\xi_1$ and $\xi_2$, and using (2.11) and (2.12), since $Y$ is non-stochastic, it follows that

$$\text{var}_{\xi_1 \xi_2}^1 (Y^*) = \left[ (U \otimes I_M) \left( \bigoplus_{x=1}^N U^{(x)} \right) \right] \text{var}_{\xi_1 \xi_2} (W) \left[ (U \otimes I_M) \left( \bigoplus_{x=1}^N U^{(x)} \right) \right]'$$

Now,

$$\text{var}_{\xi_1 \xi_2} (W) = D_{r^2} = \bigoplus_{x=1}^N D_{r^2},$$

where

$$r^2 = \left( r_1^2' \ | \ r_2^2' \ | \ \cdots \ | \ r_N^2' \right)'$$

and $D_a$ represents a diagonal matrix with diagonal elements equal to the elements of $a$. Thus,

$$\text{var}_{\xi_1 \xi_2}^1 (Y^*) = \left[ (U \otimes I_M) \left( \bigoplus_{x=1}^N U^{(x)} \right) \right] D_{r^2} \left[ (U \otimes I_M) \left( \bigoplus_{x=1}^N U^{(x)} \right) \right]'$$

Since $E_{\xi_1 \xi_2} (Y^*) = Y$,

$$\text{var}_{\xi_1 \xi_2} (Y^*) = E_{\xi_1 \xi_2} \left[ \left[ (U \otimes I_M) \left( \bigoplus_{x=1}^N U^{(x)} \right) \right] D_{r^2} \left[ (U \otimes I_M) \left( \bigoplus_{x=1}^N U^{(x)} \right) \right]' \right] + \text{var}_{\xi_1 \xi_2} [Y], \quad (A.1)$$

with $Y$ defined by (2.10).

Using the conditional expansion, we express the first term in (A.1) as

$$E_{\xi_1} \left[ (U \otimes I_M) E_{\xi_2} \left[ \left( \bigoplus_{x=1}^N U^{(x)} \right) D_{r^2} \left( \bigoplus_{x=1}^N U^{(x)} \right)' \right] \right] \left( U \otimes I_M \right)'$$

Then

$$E_{\xi_2} \left[ \left( \bigoplus_{x=1}^N U^{(x)} \right) D_{r^2} \left( \bigoplus_{x=1}^N U^{(x)} \right)' \right] = \bigoplus_{x=1}^N E_{\xi_2} \left[ U^{(x)} D_{r^2} U^{(x)}' \right].$$

The matrix in parentheses is a diagonal matrix with diagonal elements interchanged for all positions on the diagonal.
As a result, defining $\bar{\sigma}_v^2 = \sum_{i=1}^{M} \frac{\sigma_i^2}{M}$, the expectation is given by $E_{\xi_i|\xi_j} \left[ U^{(s)} D_{\xi} U^{(s)^T} \right] = \bar{\sigma}_v^2 I_M$.

Thus,

$$E_{\xi_i|\xi_j} \left[ \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^{N} U^{(s)} \right) \right] D_{\xi_i} \left[ \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^{N} U^{(s)^T} \right) \right] = E_{\xi_i} \left[ \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^{N} \bar{\sigma}_v^2 I_M \right) \left( U \otimes I_M \right)^T \right]
= E_{\xi_i} \left( UD_{\bar{\sigma}_v^2} U^T \right) \otimes I_M$$

where $D_{\bar{\sigma}_v^2}$ is an $N \times N$ diagonal matrix with elements $\bar{\sigma}_v^2$ on the diagonal. In a similar manner, we find that $E_{\xi_i} \left( UD_{\bar{\sigma}_v^2} U^T \right) \otimes I_M = \sigma_i^2 I_{NM}$, where $\sigma_i^2 = \sum_{s=1}^{N} \frac{\sigma_i^2}{N}$. As a result,

$$E_{\xi_i|\xi_j} \left[ \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^{N} U^{(s)} \right) \right] D_{\xi_i} \left[ \left( U \otimes I_M \right) \left( \bigoplus_{s=1}^{N} U^{(s)^T} \right) \right] = \sigma_i^2 I_{NM} \quad \text{(A.2)}$$

We evaluate $\text{var}_{\xi_i} (Y) = \text{var}_{\xi_i} \left[ E_{\xi_j|\xi_i} (Y) \right] + E_{\xi_i} \left[ \text{var}_{\xi_j} (Y) \right]$ next. The first term in this expansion is given by $\text{var}_{\xi_i} \left[ E_{\xi_j|\xi_i} (Y) \right] = \text{var}_{\xi_i} (X \mu + Z \beta)$. Now

$$\text{var}_{\xi_i} \left[ E_{\xi_j|\xi_i} (Y) \right] = \text{var}_{\xi_i} (U \beta) \otimes J_M \quad \text{var}_{\xi_i} (U \beta) \text{ is equal to the variance of a random permutation of cluster latent values in the population. As a result,}
\text{var}_{\xi_i} \left[ E_{\xi_j|\xi_i} (Y) \right] = \sigma^2 (P_N \otimes J_M) \quad \text{where } P_a = I_a - \frac{J_a}{a} \quad \text{We evaluate the second term in the conditional expansion of the variance in a similar manner. For the cluster in position}
\text{i in a given permutation of clusters, the variance of the SSUs is } \text{var}_{\xi_j|\xi_i} \left( Y_i \right) = s_i^2 P_M
\text{where } s_i^2 \text{ is the realization of the random variable } S_i^2 = \sum_{s=1}^{N} U_{is} \sigma_i^2 . \text{ Also, due to}
independence of the permutations of units, \( \text{cov}_{\xi_i \xi_{i'}} \left( Y_i, Y_{i'} \right) = 0 \) for \( i \neq i' \). Combining these results for all \( i = 1, \ldots, N \), \( \text{var}_{\xi_i \xi_{i'}} \left( Y_i \right) = D_s \otimes P_M \) 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 clusters, \( E_{\xi_i} \left( S_i^2 \right) = \sum_{s=1}^{\text{N}} \frac{\sigma_i^2}{N} = \sigma^2 \), and hence

\[
\text{var}_{\xi_i \xi_{i'}} \left( Y \right) = \sigma^2 \left( P_N \otimes J_M \right) + \sigma^2 \left( I_N \otimes P_M \right). \tag{A.3}
\]

Adding (A.2) to (A.3) results in (2.21).
APPENDIX B. DEVELOPMENT OF THE EXPECTED MSE

We develop the expected MSE for the predictor given by equation (3.2) of the random variable $T = (g'_{I}, g'_{II})\begin{pmatrix} Y_{I} \\ Y_{II} \end{pmatrix}$. First, we express the predictor as

$$\hat{T} = \left(g'_{I}A' + g'_{II}B'\right)Y'_{I}$$

where $A' = \begin{bmatrix} J_{nm} \otimes \frac{J_{m}}{m} + (1 - \rho_{r})k^{*} \left(\frac{P_{n}}{m} \otimes \frac{J_{m}}{m}\right) + \rho_{r}P_{nm} \end{bmatrix}$ and

$$B' = \frac{(NM-nm)\otimes \rho_{m}}{nm} + k^{*} \begin{bmatrix} \frac{J_{m}}{n} \otimes \frac{(M-m)\otimes \rho_{m}}{m} \end{bmatrix}.$$ As a result,

$$\hat{T} - T = \begin{pmatrix} Y'_{I} \\ Y'_{II} \end{pmatrix} - \begin{pmatrix} g'_{I} \\ g'_{II} \end{pmatrix} \begin{pmatrix} Y_{I} \\ Y_{II} \end{pmatrix}.$$ Using the expression for

$$\text{var}_{\hat{T}} \begin{pmatrix} Y'_{I} \\ Y'_{II} \end{pmatrix} = \begin{pmatrix} V'_{I} \ & \ & V_{I,II} \\ \ & V_{I,II} \ & \ & \ \ & \ & V_{II,II} \end{pmatrix}$$

the MSE which simplifies to

$$MSE(\hat{T}) = \begin{pmatrix} g'_{I} \\ 1 \ & \ & 1 \end{pmatrix} \begin{pmatrix} C_{11} & C'_{21} \\ \ & C_{21} \ &= \ & C_{22} \end{pmatrix} \begin{pmatrix} g'_{I} \\ 1 \ & \ & 1 \end{pmatrix}$$ (B.1)

where $C_{11} = (A' - I_{nm})V_{I}(A - I_{nm}) + \sigma_{r}^{2}A'A$, $C_{21} = (B'V_{I} - V_{II,II})(A - I_{nm}) + \sigma_{r}^{2}B'A$, and $C_{22} = B'V_{I}B - 2V_{II,II}B + V_{II} + \sigma_{r}^{2}B'B$.

Simplifying terms, $C_{11} = \left[\rho_{r}I_{nm} + (1 - \rho_{r})k^{*} \left(I_{n} \otimes \frac{J_{m}}{m}\right) - (1 - \rho_{r})(k^{*} - 1)\frac{J_{nm}}{nm}\right]\sigma_{r}^{2}$.
\[
C_{21} = \begin{pmatrix}
\sigma_r^2 & k^* \left( I_n \otimes \frac{J}{m} \right) - (k^*-1) \frac{J}{nm} \\
\frac{J}{(N-n)M \times nm} & \frac{J}{nm}
\end{pmatrix}
\sigma_r^2, \text{ and}
\]

\[
C_{22} = \begin{pmatrix}
\sigma_e^2 I_{n(M-m)} + k^* \left( \sigma_e^2 + \sigma_r^2 \right) \left( P_n \otimes \frac{J}{m} \right) & \left( \sigma_e^2 + \sigma_r^2 \right) \frac{J}{nm} \\
\left( \sigma_e^2 + \sigma_r^2 \right) \frac{J}{nm} & \left( \sigma_e^2 + \sigma_r^2 \right) \frac{J}{nm} + \sigma_e^2 I_{(N-n)M} + m \sigma_e^2 \left( I_{N-n} \otimes \frac{J}{m} \right)
\end{pmatrix}
\]

The expected MSE given by (3.7) is obtained by setting \( g' \) equal to (2.17) in (B.1) and simplifying. Expressions for the expected MSE of the predictor of \( T_A \) given by (3.4) can be developed in a similar manner.
REFERENCES

Figure 1. Percent Increase in Expected MSE for Mixed Model (——) and Scott and Smith Model (- - - ) Predictors Relative to the Random Permutation Model Predictors of the Latent Value of a Realized Sample PSU by SSU Sampling Fraction (f=1,.5,.9) and Cluster Intra-class Cluster Correlation ( \( \rho_i = 0.1,0.5,0.9 \)).

* Assuming N=100, n=30, M=20, and homogeneous unit and response error variances.