Comparison of the performance of best linear unbiased predictors (BLUP)

Peikang Yao
Synthes Spine
1302 Wrights Lane East
West Chester, PA 19380 USA
yao.peter@synthes.com

Edward J. Stanek III
Department of Public Health
401 Arnold House
University of Massachusetts
711 North Pleasant Street
Amherst, MA 01003-9304 USA
stanek@schoolph.umass.edu

ABSTRACT

The best linear unbiased predictors (the Cluster mean, the Mixed model, Scott & Smith's predictor and the Random Permutation model) of selected important public health variables were evaluated in practical settings via simulation studies. The variables corresponded to measures of diet, physical activity, and other biological measures. The simulation evaluated and compared the mean square errors (MSE) of those four predictors. It estimated variances between subjects and days, and response errors for parameters defined over one year period, based on data from a large-scale longitudinal study, the Season Study. Then, it evaluated the relative MSE increase between predictors of the true subject's mean in various settings based on theoretical results. In addition, a simulation compared the theoretical and the simulated MSE for all four predictors. The difference in the MSE between predictors was illustrated in 2D plots.

Contact Address:
Peikang Yao
Synthes Spine
1302 Wrights Lane East
West Chester, PA 19380 USA
Phone: 610-719-5283
Fax: 610-719-5102
yao.peter@synthes.com

Keywords: super-population, best linear unbiased predictor, two-stage sampling, random permutation

1. INTRODUCTION

Many clinical studies in health science field have measures of biological and behavioral variables on patients at the baseline, and then by the follow up visits. The measures fluctuate over time due to seasonal changes. In order to estimate the mean daily saturated fat intake for a subject, we usually measure total saturated fat intake on a subject on some sampled days, and then average those measurements. We would expect the estimated average based on those measurements, $\bar{X}_s$, to differ from the true subject mean, m_s , representing the mean of the true saturated fat intake over all days in a year for the subject. We call this mean, m_s , the subject's latent value. Therefore, there is a deviation between the estimated subject mean, $\bar{X}_s$, and the true subject mean, m_s . This paper is a study of predictors of the true sampled subject's mean total saturated fat intake over the whole year, in a setting where a random sample of subjects is selected from a finite population. The purpose of this paper is to discuss the extent to which best linear unbiased predictor (The Cluster mean, the Mixed model, Scott & Smith's predictor and the Random Permutation model) provides better predictor.

We evaluate properties of the predictors for different settings in the context of a large observational longitudinal study of seasonal variation in cholesterol levels which we refer to as the Season study. This study investigated the nature and causes of seasonal fluctuations in blood cholesterol. Our focus was not on seasonal effects, but rather on estimators of patient parameters such as the mean total saturated fat intake over the whole year.

We begin with a brief review of predictors that have been proposed in this setting, including the mixed model, Scott and Smith's Predictor, and the random permutation predictor. Then we illustrate the differences in interpretation, in the shrinkage constants, and in the expected MSE. Next we describe a variety of situations where prediction of key public health variables is of interest. We consider variables on diet intake, physical activity, and cholesterol and blood pressure. For those variables, we describe the accuracy of the predictors with known variances, and the results of a simulation study in the practical setting where variances have to be estimated.

2. BEST LINEAR UNBIASED PREDICTORS

Best Linear Unbiased Predictors (BLUP) were first developed by Goldberger in 1962. His emphasis was on prediction of a future observation based on past observations using estimated random effects (Goldberger, 1962). Since then, BLUP have been developed to predict unobserved values of random variables from observed values of those random variables in the sample (Scott and Smith, 1969; Royall, 1976; Robinson, 1991; Searle, Casella and McCulloch, 1992; Stanek and Singer, 2004). BLUP predictors satisfy the following conditions. First, the predictor is a linear function of the sample values. These sample values are realized random variables in the sample. Second, the predictor is required to be unbiased. Third, the predictor is required to have minimum variance.

2.1 Mixed Models

A simple mixed model for the response of the j^{th} selected unit, $j = 1,...,m$ for the i^{th} selected cluster (PSU, primary sampling unit), $i = 1,...,n$ is given by

$$Y_{ij} = \mu + B_i + E_{ij} \tag{2.1}$$

where μ denotes to the expected response over clusters in a population, B_i corresponds to the deviation between μ and the expected response of the i^{th} PSU, which is a random effect, and E_{ij} represents a random deviation of the j^{th} units' response from the expected response of the i^{th} PSU. The assumptions are that $B_i :\ iid\ N\left(0\ ,\ \sigma^2\right)$ and $E_{ij} :\ iid\ N\left(0\ ,\ \sigma_i^2\right)$ (Searle and McCulloch, 1992).

In model (2.1), we assume that variances are known, and the estimator of the fixed effect is the weighted least square estimator, $\hat{m} = \overset{n}{\underset{i=1}{\text{å}}} w_i\bar{Y}_i$ where $w_i = \dfrac{1/v_i}{\sum_{i*=1}^{n} 1/v_{i*}}$, $\bar{Y}_i = \dfrac{1}{m}\overset{m}{\underset{j=1}{\text{å}}} Y_{ij}$, and

$v_i = \sigma^2 + \dfrac{\sigma_i^2}{m_i}$. Solution of Henderson’s mixed model equations results in the BLUP of B_i. The expression $\hat{B}_i$ is given by

$$\hat{B}_i = \mathrm{k}(\bar{\mathrm{Y}}_\mathrm{i} - \hat{m}) \tag{2.2}$$

where $k = \dfrac{s^2}{s^2 + s_e^2/m}$, the variance ($s^2$) is the variance between clusters, i.e. $\mathrm{var}(B_i) = s^2$, and the variance (s_e^2) is variance within cluster, i.e. $\mathrm{var}(E_{ij}) = s_e^2$, the parameter m is the number of sampled units per sampled cluster. Then the predictor of the latent value of the i^{th} realized PSU $\left(\hat{p}_i = \hat{m} + \hat{B}_i\right)$ is a linear combination of m and the predictor of B_i given by

$$\hat{p}_i = \hat{m} + k(\bar{Y}_i - \hat{m}) \tag{2.3}$$

(Searle and McCulloch, 1992 and Stanek and Singer, 2004).

2.2 Scott and Smith’s Predictor

Scott and Smith (1969) proposed a two-stage sampling model to predict linear combinations of elements of a finite population from a super-population model. The finite population consisted of N clusters, each with M_i units. At the first stage, n clusters were selected from N clusters. At the second stage, *m* distinct units were selected from each of the n selected clusters. They assumed that:

(1) The M_i elements, $Y_{ij}s$ in the i^{th} cluster, were independent observations from a distribution with mean m_i and variance σ_i^2.

(2) The cluster means μ_1, …, m_N were uncorrelated from a distribution with mean m and variance δ^2.

The assumptions lead to a multivariate distribution for the elements with $E(Y_{ij}) = \mu$ and

$$\begin{aligned} \text{cov}\left(Y_{ij}, Y_{kl}\right) &= \delta^2 + \sigma_i^2 \text{ when } i = k; j = l \\ &= \delta^2 \text{ when } i = k; j \neq l \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.4}$$

where d^2 is variance between clusters, and s_i^2 is variance within a clusters. Scott and Smith assumed the population was a realization of the super-population, and a predictor became a linear function of the finite population units. Based on minimizing the expected MSE of a linear predictor, they developed the predictor

$$\hat{p}_i = \frac{m}{M}\overline{Y}_i + \frac{M-m}{M}[\hat{\mu}* + k_i^*(\overline{Y}_i - \hat{\mu}*)] \tag{2.5}$$

when PSU *i* is in the sample, where

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

This predictor is for PSU *i* in the sample. If the PSU is not in the sample, the mean is predicted by $\hat{p}_i = \hat{\mu}*$. The first term in the predictor (2.5) is the sample mean for the i^{th} PSU in the sample. The second term is the predictor of the remaining secondary sampling units (SSUs) for the PSU. The weight factors were the ratio of observed SSUs and the unobserved SSUs. If PSUs were not in the sample, the predictor simplified to be the weighted sample mean (Stanek and Singer, 2004).

2.3. The Finite Population Response Error Model and Parameterizations

A third predictor was developed based on a two stage random permutation of the population (Stanek and Singer, 2004). Assume that a finite population is composed of *N* clusters, indexed by $s=1,...,N$, and each cluster contains a listing of *M* units, indexed by $t=1,...,M$. Assume the k^{th} response for unit t in cluster s is given by

$$Y_{stk}=y_{st}+W_{stk}\,. \tag{2.6}$$

where y_{st} donates a fixed constant representing the expected response for the unit, and W_{stk} represents response error (with zero expected value). Model (2.6) is referred to as a response error model.

If σ_{st}^2 represents the response error variance for unit t in cluster s, then the average response error variance is given by $\sigma_r^2=\sum_{s=1}^{N}\sum_{t=1}^{M}\frac{\sigma_{st}^2}{NM}$. The mean and the variance of the expected response for units in cluster s are defined as $\mu_s=\frac{1}{M}\sum_{t=1}^{M}y_{st}$ and $\left(\frac{M-1}{M}\right)\sigma_s^2=\frac{1}{M}\sum_{t=1}^{M}\left(y_{st}-\mu_s\right)^2$ for $s=1,\ ...,\ N$, respectively. The average within cluster variance is defined as $\sigma_e^2=\frac{1}{N}\sum_{s=1}^{N}\sigma_s^2$.

Similarly, the population mean and the variance between clusters can be defined as $\mu=\frac{1}{N}\sum_{s=1}^{N}\mu_s$ and $\left(\frac{N-1}{N}\right)\sigma^2=\frac{1}{N}\sum_{s=1}^{N}\left(\mu_s-\mu\right)^2$, respectively. The deviation of the latent value of cluster s, μ_s, from the population mean is represented as $\beta_s=\left(\mu_s-\mu\right);$ the deviation of the expected response for unit t (in cluster s) from the latent value of cluster s is donated as $\varepsilon_{st}=\left(y_{st}-\mu_s\right)$. Then, the response for unit t in cluster s is given by

$$y_{st}=\mu+\beta_s+\varepsilon_{st}\,. \tag{2.7}$$

Assuming $\mathbf{y}=\left(\mathbf{y}_1{}' \quad \mathbf{y}_2{}' \quad ... \quad \mathbf{y}_N{}'\right)'$ where $\mathbf{y}_s=\left(y_{s1} \quad y_{s2} \quad ... \quad y_{sM}\right)'$, model (2.7) can be summarized for all units as

$$\mathbf{y} = \mathbf{X}\mu + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.8}$$

where $\mathbf{X} = \mathbf{1}_N \otimes \mathbf{1}_M$, $\mathbf{Z} = \mathbf{I}_N \otimes \mathbf{1}_M$, $b' = (b_1\ b_2\ \ldots\ b_{\mathrm{N}})$. None of the terms in model (2.8) are random variables. $\mathbf{1}_a$ is an $a \times 1$ column vector of ones, and $\boldsymbol{\varepsilon}$ is defined similarly to $\mathbf{y}$.

2.4 Two Stage Random Permutation Model

Let us index a cluster in a permutation by $i = 1,\ 2,\ \ldots, N$. A random variable U_{is} is one when PSU *i* is cluster s, zero otherwise. We define the units in a cluster as secondary sampling units (SSU). Similarly, a random variable $U_{jt}^{(s)}$ takes value of one when SSU j in cluster *s* is unit *t*, and zero otherwise. When all permutation lists are equally likely, the random vector $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathrm{N}})'$ is a random permutation of the population, and each element of $\mathbf{Y}_i = (\mathrm{Y}_{i1}\ \mathrm{Y}_{\mathrm{i}2}\ \mathrm{Y}_{\mathrm{i}3}\ \ldots\ \mathrm{Y}_{\mathrm{iM}})'$. The random variable representing the j^{th} SSU within the i^{th} PSU, Y_{ij} , in the permutation is as follows:

$$Y_{ij} = \sum_{s=1}^{N} \sum_{i=1}^{M} U_{is} U_{jt}^{(s)} Y_{st} \tag{2.9}$$

where U_{is} takes a value of one when PSU i is cluster s, and a value of zero otherwise, and U_{jt} takes on a value of one when SSU j in cluster *s* is unit *t*, and zero otherwise.

We assume a total of m elements in each of n clusters are selected by a two-stage sampling scheme from a population. The population total is composed of three components: 1) the total for observed elements; 2) the total for unobserved elements in sampled clusters; and 3) the total for unsampled clusters. Stanek and Singer (2004) developed an unbiased predictor of a realized PSU mean, which was a linear combination of the random variables in the sample. The predictor minimized the expected values of the mean square error (MSE). If there was no response error, the mean of PSU *i* can be predicted by

$$\hat{T}_i = f\bar{Y}_i + (1-f)\left(\bar{Y} + k\left(\bar{Y}_i - \bar{Y}\right)\right) \tag{2.10}$$

where $f = \dfrac{m}{M}$ (sampled fraction), $\bar{Y}_i$ is the mean of i^{th} selected cluster, $\bar{Y}$ is the overall mean

defined as $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^{n} \bar{Y}_i$, $k = \dfrac{m\sigma^{*2}}{m\sigma^{*2} + \sigma_e^2}$ and $\sigma^{*2} = \sigma^2 - \dfrac{\sigma_e^2}{M}$.

2.5 Comparison Between Predictors

Table 1 presents the predictors illustrated previously. The predictors appears to be algebraically similar, but their contents are different. Each predictor is a weighted linear combination of two terms. The first term predicts the latent values of the SSUs in the sample. The second term predicts the latent values for the remaining SSUs not in the sample. Unlike other predictors, the mixed model predictor is a predictor of the unobserved SSUs for a PSU, and it places all the weight on the second term.

Table 1. Predictors of the latent value of PSU i when $i \le n$ in two-stage cluster sampling (Stanek, 2003)

Model	**Predictor**	
Cluster Mean	$\hat{P}_i = f\bar{Y}_i$	$+(1-f)\bar{Y}_i$
Mixed Model	$\hat{p}_i =$	$\left(\hat{\mu} + k_i\left(\bar{Y}_i - \hat{\mu}\right)\right)$
Scott & Smith	$\hat{P}_i = f\bar{Y}_i$	$+(1-f)\left(\hat{\mu}^* + k_i^*\left(\bar{Y}_i - \hat{\mu}^*\right)\right)$
Random Perm.	$\hat{T}_i = f\bar{Y}_i$	$+(1-f)\left(\bar{Y} + k\left(\bar{Y}_i - \bar{Y}\right)\right)$
RP + Resp. Err.	$\hat{T}_i = f\left(\bar{Y} + k_r^*\left(\bar{Y}_i - \bar{Y}\right)\right)$	$+(1-f)\left(\bar{Y} + k^*\left(\bar{Y}_i - \bar{Y}\right)\right)$

In these expressions, $\bar{Y}_i = \frac{1}{m}\sum_{j=1}^{m} Y_{ij}$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_i$, $\hat{\mu} = \sum_{i=1}^{n} w_i \bar{Y}_i$, $w_i = \frac{1/v_i}{\sum_{i^*=1}^{n} 1/v_{i^*}}$,

$v_i = \sigma^2 + \frac{\sigma_i^2}{m}$, $\hat{\mu}^* = \sum_{i=1}^{n} w_i^* \bar{Y}_i$; $w_i^* = \frac{1/v_i^*}{\sum_{i^*=1}^{n} 1/v_{i^*}^*}$, $v_i^* = \delta^2 + \frac{\sigma_i^2}{m}$,

$k_i = \frac{m\sigma^2}{m\sigma^2 + \sigma_i^2}$, $k_i^* = \frac{m\delta^2}{m\delta^2 + \sigma_i^2}$, $k_e = \frac{m\sigma^{*2}}{m\sigma^{*2} + \sigma_e{}^2}$, $k_r^* = \frac{m\sigma^{*2} + \sigma_e^2}{\left(m\sigma^{*2} + \sigma_e^2\right) + \sigma_r^2}$, and

$k^* = \frac{m\sigma^{*2}}{m\sigma^{*2} + \left(\sigma_e^2 + \sigma_r^2\right)}$, $s^{*2} = s^2 - \frac{s_e{}^2}{m}$. Note that in the Mixed Model, or in Scott and Smith's Predictor without response errors and assuming equal unit variances for each cluster, the expression for σ_i^2 is equal to $\left(\sigma_e^2 + \sigma_r^2\right)$ (Stanek, 2003).

Scott and Smith's predictor and the random permutation model are nearly identical. However, the differences between the predictors is due to differences in variance components and shrinkage constants (Stanek and Singer, 2004). Because the random permutation model permuted PSUs, it uses a single SSU component of the variance representing the average of the SSU within cluster variance. Therefore, the variance within a cluster is s_e^2 (as supposed to the cluster specific components, s_i^2, in the Scott and Smith's predictor). Similarly, the variance between PSU in the random permutation model is defined as $\sigma^{*2} = \sigma^2 - \frac{\sigma_e^2}{M}$, as opposed to d^2 in Scott and Smith's predictor. We note that the assumption of Scott and Smith's predictor for the variance components does not correspond to the variance components that derived from permutation clusters and units in a finite population (Stanek and Singer, 2004).

With additional assumptions that the variance within a cluster is identical for all clusters, s_e^2 and the response error variance is the same for all units and equal to s_r^2, and $\sigma_i^2 = \sigma_e^2 + \sigma_r^2$, then the shrinkage constants $k_i = k_i^* = \frac{m\sigma^2}{m\sigma^2 + \sigma_e^2 + \sigma_r^2}$. Each predictor in Table 1 can be represented as $\hat{T} = \bar{Y} + c\left(\bar{Y}_i - \bar{Y}\right)$. Table 2 shows values of c for Predictors $\hat{T} = \bar{Y} + c\left(\bar{Y}_i - \bar{Y}\right)$ of the Latent Value of PSU i when $i \le n$ in Two-stage Cluster Sampling with Homogeneous Unit and Response Error Variances. As shown in Table 2, the differences between the predictors are due to the contents of shrinkage constants.

Table 2. Values of c for predictors $\hat{T} = \bar{Y} + c\left(\bar{Y}_i - \bar{Y}\right)$ of the latent value of PSU i

Model	
Mixed Model	$c_{MM} = k_i$
Scott & Smith	$c_{SS} = f + (1-f)k_i$
Random Permutation.	$c_{RP} = f + (1-f)k$
Random Permutation with Response Error	$c_{RPR} = f\rho_t + (1-f\rho_t)k^*$

2.6 Simulation

A simulation study program was developed by Stanek (2003a) to evaluate the predictors in two-stage cluster sampling contexts in different settings. The simulation study consisted of three sub-modules, creation of a finite population, selection of two stage samples from the finite population, and comparison of the simulated mean square errors (SMSE) and the theoretical mean square errors (TMSE) of the predictors.

First, the population was defined by a set of values based on percentiles of a hypothetical distribution. We used percentiles of such distribution to create a population of units and clusters. The basic distributions from which the finite populations were generated were normal. Different distributions could be selected for units and clusters. However, we used the same distribution (normal) to generate the unit effect for all clusters.

For each simulation, the population consisted of N clusters with M units per cluster. Each individual cluster parameter was represented by the latent value for cluster *s*, μ_s and their mean, by μ. The variance between clusters, σ^2, was fixed, and then N initial values evenly spaced were generated based on the percentiles of the specified distribution. These values were the initial values of the cluster parameters. Because the number of clusters in the population was finite, the average of the cluster parameters would not necessarily equal to be the population mean, μ. The mean of cluster parameters was redefined by centering them at μ and re-scaling their values so that the variance matched $\sigma^2 = \sum_{s=1}^{N} \frac{(\mu_s - \mu)^2}{N-1}$.

Next, we generated unit effects for the M units for each cluster. These effects were to be forced to average to zero. In addition, the unit effects were generated using percentiles of a specified distribution. The distribution was normal. The variance of the unit effects may be either set to be constant for all clusters or vary proportionally to $\sqrt{\max\left(0.1,\ \frac{\mu_s}{\mu}\right)}$. The parameters for the cluster were formed by adding the unit effect to the cluster mean and were represented by y_{st}. The variance of the unit parameters in cluster s was given by $\sigma_s^2 \ = \ \sum_{t=1}^{M} \frac{(y_{st} - \mu_s)^2}{M-1}$. The common within cluster variance, represented by σ_e^2, was equal to the average within cluster variance, $\sigma_e^2 = \sum_{s=1}^{N} \frac{\sigma_s^2}{N}$. Unit effects were re-scaled so that they had zero mean for each cluster, and their average variance was equal to be σ_e^2.

The parameters in the simulation programs were the number of clusters in the population (N), the number of units in a cluster (M), the number of clusters in the sample (n), the number of units in a sampled cluster (m), the population mean (μ), the variance of cluster means (σ^2), the variance of units in a cluster (σ_s^2), the response error variances (s_r^2), the cluster distribution, and the unit distribution.

We adopted and modified this simulation for this research. The average differences between the predicted PSU mean and the true PSU mean, and the MSE were estimated based on known and unknown variances. If variance components were known, there were eight different assumptions about mixed model, leading to eight mixed model predictors, and eight analogous predictors based on Scott & Smith’s model. Also we obtained the predictor corresponding to the cluster mean, and to the Random Permutation model. This resulted in total to 18 predictors based on known variances. For each predictor, we estimated the MSE by evaluate the average squared deviation from the realized cluster mean over many samples (where each sample corresponds to a trial). If variance components were unknown, Stanek (2003a) proposed two different methods of estimating variance components, which lead to two mixed model predictors, two Scott and Smith’s predictors, and two random permutation predictors.

3. MATERIALS AND METHODS

3.1 Season Study

Data sets for this project were from the Season Study and provided by Dr. Ockene. The variables included cholesterol, diet, light, activity and other variables with either single or multiple measurements on each subject. Data were collected from volunteers (N=5000) recruited from Fallon Health Maintenance Organization (HMO) members with age between 20 and 70 year old. Patients were enrolled between 1994 and 1998. Measurements were conducted on consecutive three-month intervals over a twelve-month period on each subject (Stanek and Singer, 2004). The data sets consisted of three sub-data sets: quarterly data, 24-hour recall data, and baseline data.

The Quarterly data contained lipid data, 7DDR (7 day dietary records based on subject recall, 1 each per quarter), and hormonal data. Measurements were made on a simple random sample of 1 day ($m = 1$) during a quarter (3 month period, $M = 90$). The 24-hour recall data contained physical activity, diet, and light exposure. They were measured on 2 randomly selected weekdays and one randomly selected weekend day per quarter $(m = 3,\ M = 90)$. Nutrient variables evaluated were total saturated fat intake, total fat, total carbohydrate, and cholesterol. Physical activity variables were measured to evaluate the 1-year average of reported physical activity energy expenditure (MET-hours/day), by activity domain (i.e., household, occupational and leisure time) and intensity. Standard metabolic equivalent (MET) values were used to calculate estimates of physical activity energy expenditure. A weighted sum of daily physical activity energy expenditure (MET-hours/day) was computed using the time reported (hours/day) in activity of each intensity and the following MET weights: light activity, 1.5 METs; moderate activity, 4.0 METs; vigorous activity, 6.0 METs; and very vigorous activity, 8.0 METs. One MET-hour/day was approximately equivalent to 1 kcal/kg body mass/hour or to the resting metabolic rate of a person weighting 60-70 kg (Matthews, et al., 2001). Physical activity variables to be evaluated were light intensity activity, moderate intensity activity, and vigorous intensity activity. Up to 3 days of 24-hour activity were collected per quarter. The baseline data included demographic factors, such as age and gender.

Both subjects and dates of measurements over a quarter were not randomly selected. Subjects participating in this study were assumed to be comparable to a simple random selection from a finite population. Therefore, two-stage Best Linear Unbiased Predictors in this study were assumed to be applicable for this dataset.

3.2 Estimating Variance Components

Variance components such as variance between clusters (subjects), residual variance (a combination of variance between units (days) and response error) for one year (i.e. M=365) were estimated using a mixed model with restricted maximum likelihood method using the Season's study data. The variance between days (σ_d^2) was calculated by subtracting response error (σ_e^2) from the residual variance. The intra-class correlation of repeated measures on a unit (ρ_t), and the intra-class correlation of units in a cluster (ρ_s) were estimated based on the variance components $\left(i.e.\ \rho_t = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_e^2},\ \rho_s = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2} \right)$. Response errors of variables are based either on literature reviews or through simulations with coefficient variations (Table 4). For example, Hegsted and Nicolosi (1987) estimated response error for total serum cholesterol was 225 (mg/dl).

If response errors of the variables were not available in the literature, they were estimated by simulating two responses on each subject. First, response error of a measure on each subject was calculated based on assumed coefficient of variation. Then, the average response error was estimated by pooling individual response error on each subject. Two simulated responses were based on empirical measures of each subject at the quarter 1 of the quarterly data set (cbdq6) in the Seasons Study and using the coefficient of variations assumed (Table 4).

We simulated two responses of weight for each subject with random normal function generator given by

$$Y_{sdk} = \mu_{sd} + \mu_{sd}\text{*c.v.*rannor(seed)} \qquad (3.1)$$

where Y_{sdk} represents weight of subject s at the day *d* and at the k^{th} measure, μ_{sd} denotes the true weight of subject s on a day *d*, which is assumed to be the mean weight of subjects at quarter 1 of quarterly data in Season Study, i.e. 78.44 (kg), the term *rannor(seed)* is a random number function based on a normal distribution. $\mu_{sd}\text{*c.v.}$ is the possible response error allowed to be different for different subjects and days. The response error of measures in weight was estimated using a mixed model (see Appendix A).

A simple mixed model, which was used to estimate variance components for a variable, for the response of the j^{th} selected day, j = 1,...,m, for the i^{th} selected subject, i = 1,...,n, is given by

$$Y_{ijk} = m + \mathrm{B}_{\mathrm{i}} + \mathrm{D}_{\mathrm{ij}} + \mathrm{E}_{\mathrm{ijk}} \tag{3.2}$$

where μ denotes to the expected response over subjects in a population, B_i corresponds to the deviation between μ and the expected response of the i^{th} selected subject. D_{ij} indicates the deviation between expected response of the i^{th} selected subject at the j^{th} selected day, Y_{ij}, and the expected response of subject i, and E_{ijk} represents the random response error of measures for the i^{th} selected subject and at the j^{th} selected day. B_i, D_{ij} and $\mathrm{E}_{\mathrm{ijk}}$ are random. The assumptions are that $B_i :\ iid\ N\left(0\ ,\ \sigma^2\right)$.

Variables chosen to estimate variance components using the mixed model (3.2) were body composition and lipid variables (body mass index, systolic blood pressure, diastolic blood pressure, LDL, HDL, total cholesterol, triglyceride, weight), nutrient variables (total saturated fat intake, total fat intake, saturated fat as percent of total calories, total fat as percent of total calories) and physical activity (light intensity activity, moderate intensity activity and vigorous intensity activity). The quarterly data contained lipid data, collected by one measure per subject per quarter, and nutrient data, collected by 7 day dietary recall with 1 each per quarter, and nutrient and physical activity data, collected by 24 hour recall telephone interview given up to 3 days per quarter.

One of the assumptions to estimate variance components for three sets of variables was that a subject was selected at random. All subjects with five measures (one per quarter) were retained. The variance components of lipid, nutrient variables (variance between subjects, and the residual) with a period of 365 days were estimated with the mixed model shown in (3.2) (see Appendix B).

Similarly, 365-day-period variance components of nutrient and physical activity variables collected by 24-hour recall were estimated again with the mixed model illustrated in (3.2) (See Appendix B). Because a residual error was composed of variance between days and a response error, a 365-day-period variance between days was the deviation between the residual error and response error.

All analyses were performed using SAS 8.02.

3.3 The Simulation Study

The simulation study program addressed in Section 2.6 was modified to compare the performance of Best Linear Unbiased Predictors. Modification included calculating the relative MSE increases between predictors of the true subject mean in various settings with or without known variance components. The relative MSE is defined as $\left(\frac{mse1-mse2}{mse2}x100\%\right)$, where $mse1$ is either the theoretical or the simulated MSE of the predictors, the predictors are the Cluster mean, the Mixed model, Scott & Smith’s predictor; $mse2$ is the theoretical MSE of the Random Permutation model. In addition, modification enabled us to compare the theoretical and the simulated MSE for all four predictors with or without known variance components. Since there were eight possible ways the variance components for the Mixed model and Scott & Smith’s predictor could be defined when variance components were known, it lead to eight Mixed models and eight Scott & Smith’s predictors. The eighth mixed model and the eighth Scott & Smith’s predictor were used to compare the performance with other predictors. When variance components were unknown for those two predictors, there were two methods of estimating variance components leading to two Mixed models and two Scott & Smith’s predictors. The second Mixed model and the second Scott & Smith’s predictor were used to compare the performance with the Cluster mean model and the Random Permutation predictor.

4. RESULTS

4.1 Estimate Variance Components

Table 4 shows the parameters for coefficient of variation (c.v.) and the mean values of the variables of interest. The process of estimating the coefficient of variation and the response errors of variables is described next.

Let us consider an example of estimating intra-individual variability of physical activity level in minutes on one day for a subject. Suppose that two measures of physical activity on the same day were collected for each subject. Each subject was asked twice the number of minutes he/she exercised during the day. Suppose, for example, that two responses of the first subject are 24 and 36 minutes. Also, suppose that the actual number of minute the first subject exercised is 30 minutes. Similarly, for the second subject, the two responses are 16 and 24 minutes. The actual time the second subject exercised is 20 minutes. Then, the range (or interval) of two responses for each subject accounts for 40% of the actual values. If such data were available for n subjects, they could be used to estimate the variance components of physical activity level. Because such data are not available, we assume that a 95% confidence interval for response has a width of 40% of the actual minutes.

Then, the k^{th} response for subject s in day d , Y_{sdk} is given by

$$Y_{sdk} = m_{sd} + \mathrm{E}_{\mathrm{sdk}} \qquad (4.1)$$

where m_{sd} represents the true amount of physical activity of subject s on a day d , which is assumed to be the mean amount of physical activity of subjects at quarter 1 of quarterly data set (cbdq6) in the Seasons Study. E_{sdk} denotes the response error. One standard deviation of response errors of physical activity level of subject s at day d, (s_{e}) is $0.1\, m_{sd}$ as a result of

95% C.I. width assumption because $\sigma_e = 0.2*2*\mu_{sd}/4 = 0.1*\mu_{sd}$. Therefore, a within-subject coefficient of variation (c.v.) of physical activity level is equal to be 0.1 because $c.v. = \sigma_e/\mu_{sd}$.

Similarly, we can estimate the response errors by simulations with assumed coefficient of variance of lipid variables. Let us estimate the coefficient of variances of weight and body mass index, followed by the process of simulations.

The k^{th} response of weight for subject s, Y_{sk}, is given by

$$Y_{sk} = \mu_s + \mathrm{E}_{sk} \qquad (4.2)$$

where μ_s represents the true weight measure of subject s. E_{sk} denotes the response error of weight. Two responses of weight measures are assumed to be 0.25 kg above or 0.25 kg below the true weight. Therefore, the standard deviation of response errors of weight measures of subject s at day d is $0.25*2/4 = 0.125$. The μ_s is assumed to be equal to the mean weight of subjects at quarter 1 of quarterly data set (cbdqs6) in the Seasons Study, i.e. 78.44 kg. Then a within-subject coefficient of variation (c.v.) of response error of weight is equal to be 0.0016 because $c.v. = \sigma_e/\mu_{sd} = 0.125/78.44$.

Next, we estimate the coefficient of variance of body mass index. Suppose the true subject height (μ_s) is assumed to be 1.7 meter. Then the k^{th} response of body mass index for subject s is given by

$$\frac{Y_{sk}}{1.7^2} = \frac{\mu_s}{1.7^2} + \frac{\mathrm{E}_{sk}}{1.7^2} \qquad (4.3)$$

We assume height can be measured to the nearest centimeter (i.e. 4*standard deviation =2 cm; i.e. two height measures can be 1.69 meters and 1.71 meters). The maximum response error for body mass index is $\mathrm{var}\left(\frac{E_{sk}}{1.69^2}\right) = \left(\frac{0.25}{2.8561}\right)^2 = 0.0875^2$. Therefore, the standard deviation of response errors of body mass index measures of subject s is 0.0875. The body mass index of subject s is assumed to be the mean body mass index of subjects at quarter 1 of quarterly data set (cbdqs6) in the Seasons Study, i.e. 27.34 kg/m2. Then a within-subject coefficient of variation (c.v.) of response error of body mass index is equal to be 0.0032 because $c.v. = s_e / m_{sd} = 0.0875/27.34$.

Table 5 shows variance components of lipid, nutrient and physical variables and intra-class correlation of cluster and unit at one year. As indicated in Table 5, the variance of lipid variables such as total cholesterol between subjects for the time period of 365 days is 1463.33 mg/dl. The residual of variance components (i.e. day to day and response errors) of total cholesterol is 286.32 mg/dl. Because the response error of serum cholesterol is 225 mg/dl, the amount of total cholesterol between days for the time period of 365 days is 61.32 mg/dl. Table 5 also reports that the intra-class correlation of units in a cluster (subject), and of repeated measures on a unit (day) for total cholesterol at the time period of 365 days are 0.960 and 0.214.

In addition, Table 5 indicates that a 365-day-period variance of nutrient variables collected using 24 hour recall such as total saturated fat intake (SFA) and residual error are 113.24 gm and 169.57 gm. Because response error of total SFA is 151.01 gm, a 365-day-period variance of total SFA between days is 18.56 gm. Furthermore, Table 5 shows that the intra-class correlation of units in a cluster (subject), and the intra-class correlation of repeated measures on a unit (day) for total SFA at the time period of 365 days are 0.859 and 0.109.

Table 6 shows cluster and unit intra-class correlations of some health variables at 365-day period in tabular format. The performance of the predictors (the Mixed Model, Scott and Smith's Predictor, the Cluster Mean Model, and the Random Permutation Model) were evaluated by comparing the simulated mean square errors (SMSE) and the theoretical mean square errors (TMSE) of the models at two common settings of cluster and unit intra-class correlations ($r_s = 0.67,\ r_t = 0.83;\ r_s = 0.67,\ r_t = 0.2$).

4.2 The Simulation Results

4.2.1 Known Variances And Equal Within Cluster Variances

Figures 1 and 2 illustrate the percent increases in the difference between the simulated mean square errors (SMSE) and the theoretical mean square errors (TSME) of the predictors (the Mixed Model, Scott and Smith's Predictor, the Cluster Mean Model, and the Random Permutation Predictor) under an assumption of known variances, and with equal within-cluster variances in two common settings of cluster and unit intra-class correlations shown in Table 6 ($r_s = 0.67,\ r_t = 0.83;\ r_s\!=\!0.67, r_t\!=\!0.2$) and at two simulation runs (1000, 10000). With 1,000 trials, the relative difference is between -2.5% and 3.5%. When the unit sampling fraction (f) increases to 0.7, the relative difference reaches the peak (Figure 1, top). With 10,000 trials in both Figures 1 and 2, the relative differences have been reduced between -1.0% and 1.0%. In addition, there are no obvious peaks in 10,000 trials as occurs with 1,000 trials. Furthermore, with 10,000 trials, there are no predictors showing the consistent results with the smallest relative increment in SMSE over its TMSE through all the unit-sampling fractions.

Figure 3 shows the percent increases in the theoretical mean square errors (TMSE) of predictors (the Mixed Model, Scott and Smith's Predictor, and the Cluster Mean Model) over TMSE of the Random Permutation Model under an assumption with known variances, and with equal within-cluster variances in two common settings of cluster and unit intra-class correlation coefficients ($r_s = 0.67,\ r_t = 0.83;\ r_s\!=\!0.67, r_t\!=\!0.2$) and at one simulation (number of runs=10,000). Several patterns emerge from Figure 3 (top), where both cluster and unit intra-class correlation coefficients are larger ($r_s = 0.67,\ r_t\!=\!0.83$). First, the Random Permutation Predictor has the minimum TMSE, followed by Scott and Smith's predictor. Second, as the unit sampling fraction becomes large, the magnitude in percent increment in TMSE of the Mixed Model over TMSE of the Random Permutation Predictor increases, but the magnitude in percent increases in the TMSE of the Cluster Mean Model over the TMSE of the Random Permutation Predictor decreases. When unit-sampling fraction gets over 0.63, the Mixed Model performs worse than the Cluster Mean Model. However, several different patterns emerge from Figure 3 (bottom), where unit intra-class correlation is smaller ($r_s = 0.67, r_t = 0.2$). First, the Random Permutation Predictor has the minimum TMSE, but followed by the Mixed Model instead of Scott and Smith's Predictor. Second, as the unit-sampling fraction becomes large, the magnitude of increment in TMSE of Scott and Smith's predictor over TMSE of the Random Permutation Predictor increases gradually, but the magnitude of increment in the TMSE of the Cluster Mean Model over the TMSE of the Random Permutation Predictor decreases dramatically. When the unit sampling fraction reaches 0.9, however, the Cluster Mean Model still has higher percent increment in its TMSE over TMSE of Random Permutation Predictor than Scott and Smith's Predictor.

Figure 4 illustrates the percent increment in the simulated mean square errors (SMSE) of the predictors over the TMSE of the Random Permutation Predictor under an assumption with known variances, and with equal within-cluster variances in two common settings of cluster and unit intra-class correlations ($r_s = 0.67$, $r_t = 0.83$; r_s=0.67, r_t=0.2) and at one simulation (number of runs=10,000). There are similar patterns between Figure 4 (top) and Figures 3. For example, when the unit-sampling fraction gets over 0.6, the Cluster Mean Model performs better than the Mixed Model. However, TMSE of the Random Permutation Model is not always smaller than the SMSE of Scott and Smith's predictor as shown in Figure 3, the Cluster Mean Model performs much worse than other predictors in Figure 4 (bottom). The ratio of increment in SMSE of the Cluster Mean Model over TMSE of the Random Permutation Predictor is over 95% when unit-sampling fraction is 0.20. As the unit sampling fraction increases, the ratio gets smaller. But it is 20% when the unit sampling fraction (f) increases to 0.9.

4.2.2 Unknown Variances And Equal Within Cluster Variances

Figure 5 illustrates the percent increases in the difference between the simulated mean square errors (SMSE) and the theoretical mean square errors (TSME) of the predictors (the Mixed Model, Scott and Smith's Predictor, the Cluster Mean Model, and the Random Permutation

Model) under an assumption with unknown variances, and with equal within cluster variances in two common settings of cluster and unit intra-class correlations ($r_s = 0.67,\ r_t = 0.83;\ r_s=0.67, r_t=0.2$) respectively and at one simulation (number of runs=10,000). Several patterns emerge from Figure 5 (top). First, the percent increment in SMSE over TMSE of four predictors (excluding the Cluster Mean Model) decreases as the unit sampling fraction increases. Second, the Mixed Model has the highest increment in SMSE to TMSE in all unit-sampling fractions and over all other predictors. It reaches over 30% when the unit-sampling fraction is 0.2, and deceases to 8% when the unit-sampling fraction gets to 0.9. Among the Cluster Mean Model, Scott and Smith's Predictor, and the Random Permutation Predictor, the Cluster Mean Model has the smallest increment in SMSE over TMSE. The ratio of percent increment in SMSE to TMSE of Scott and Smith's predictor over the ratio of the Random Permutation Model is about 2 when the unit-sampling fraction is 0.2. The ratio decreases to 1 as the unit-sampling fraction gets 0.9.

Similar patterns emerge form Figure 5 (bottom) when $\rho_s = 0.67$, and $\rho_t = 0.2$. However, there are some differences. One of them is that the difference in percent increment in SMSE relative to TMSE between Scott and Smith's predictor and the Random Permutation Predictor depends on the unit-sampling fraction (f). When f is less than 0.35, Scott and Smith's predictor has higher increment in SMSE over TMSE than the Random Permutation Predictor. However, when f is greater than 0.35, the Random Permutation Predictor has higher percent increment in SMSE to TMSE than Scott and Smith's Predictor.

Figure 6 illustrates the percent increases in the simulated mean square errors (SMSE) of the predictors over the TMSE of the Random Permutation Model under an assumption with unknown variances, and with equal within-cluster variances in two common settings of cluster and unit intra-class correlations ($r_s = 0.67,\ r_t = 0.83;\ r_s=0.67, r_t=0.2$) and at one simulation (number of runs = 10,000). Several patterns appear in Figure 6 (top). First, the TMSE of Random

Permutation Predictor is less than SMSE of other predictors almost all the times. Second, Scott and Smith's predictor has the smaller percent increase in SMSE to TMSE of the Random Permutation Predictor than the Mixed Model and the Cluster Mean Model most of times. Third, the increment in SMSE of the predictors over TMSE of the Random Permutation Predictor declines as the unit sampling fraction (f) increases. However, when unit intra-class correlation (ρ_t) is 0.83 (Figure 6, top), the Mixed Model has the highest percent increase in SMSE to TMSE of the Random Permutation Predictor, followed by the Cluster Mean Model. When unit intra-class correlation (ρ_t) is 0.2 (Figure 6, bottom), the Cluster Mean Model has the highest percent increase in SMSE to TMSE of the Random Permutation Predictor, followed by the Mixed Model. Finally, when unit intra-class correlation (ρ_t) is 0.2 in Figure 6 (bottom), the percent increases in SMSE of all four predictors over TMSE of the Random Permutation model is about 175% at the unit-sampling fraction of 0.1; however, when unit intra-class correlation (ρ_t) is 0.83 in Figure 6 (top), the percent increases in SMSE of all four predictors over TMSE of the Random Permutation model is about 47%. That may imply that the higher response errors in measures of the variables, the bigger the percent increases in SMSE to TMSE of the Random Permutation model when the unit-sampling fraction is very lower such as 10%.

Table 4. The parameters for the coefficient of variation of some health science variables

Variables	Collection interval	C.V.	Mean	Source and assumptions
Total SFA	24hr	0.10	25.19 gm	A 95% confidence interval for response has a width of 40% of the true response.
	7ddr	0.20	29.02 gm	A 95% confidence interval for response has a width of 80% of the true response.
Percent of SFA	24hr	0.10	11.20 % of calories	A 95% confidence interval for response has a width of 40% of the true response.
	7ddr	0.10	12.50 of % calories	A 95% confidence interval for response has a width of 40% of the true response.
Total Fat Intake	24hr	0.10	70.01 gm	A 95% confidence interval for response has a width of 40% of the true response.
	7ddr	0.20	85.62 gm	A 95% confidence interval for response has a width of 80% of the true response.
Percent Fat Intake	24hr	0.10	31.1 % calories	A 95% confidence interval for response has a width of 40% of the true response.
	7ddr	0.10	36.91 % calories	A 95% confidence interval for response has a width of 40% of the true response.
Light intensity activity	24hr	0.10	0.21 MET hr	A 95% confidence interval for response has a width of 40% of the true response.
Moderate intensity activity	24hr	0.10	0.86 MET hr	A 95% confidence interval for response has a width of 40% of the true response.
Vigorous intensity activity	24hr	0.10	0.87 MET hr	A 95% confidence interval for response has a width of 40% of the true response.
Body weight	Quarterly	0.0016	78.44 kg	absolute difference between two responses is 0.25 kg,
Body mass index	Quarterly	0.0032	27.34 kg/m2	absolute difference between two responses is 0.25 kg, the accuracy of height measures is 0.01 meter.
TG	Quarterly	0.226	140.90 mg/dl	Smith et al, 1993.
HDL	Quarterly	0.074	47.28 mg/dl	Smith et al, 1993.
LDL	Quarterly	0.09	144.02 mg/dl	Smith et al, 1993.
SBP	Quarterly	0.03	119.83 mm Hg	Andre et al., 1987. Cavelaars et al., 2004. Ripolles et al., 2001.
DBP	Quarterly	0.032	77.15 mm Hg	Andre et al., 1987. Cavelaars et al., 2004. Ripolles et al., 2001.
TC	Quarterly	0.0688	218.58 mg/dl	Hegsted and Nicolosi (1987)

(Continued on next page)

Table 4. Continued

*24hr: Based on a 24 hours recall telephone interview given for up to 3 days in a quarter.
*7ddr: 7-day dietary records based on subject recall, 1 each per quarter.
*based on Season's study data unless otherwise noted.
Source: thesis04py_table4.sas

Table 5. Variance components of some health variables and intra-class correlation (ICC) of cluster and unit

Variables	Collection	Time	Variance components			ICC	
	Interval	Period	Subject	days	resp. error	Cluster	Unit
% FAT(%calories)	7ddr	365	39.79	10.99	15.60	0.784	0.413
% SFA(%calories)	7ddr	365	6.40	2.29	1.79	0.737	0.561
Total FAT(gm)	7ddr	365	1397.83	1037.74	401.61	0.574	0.721
Total SFA(gm)	7ddr	365	156.72	73.38	47.40	0.681	0.608
% FAT(%calories)	24hr	365	29.22	4.15	56.22	0.876	0.069
Lig. Activ.(MET hr)	24hr	365	0.02	0.00	0.57	1.000	0.000
Mod. Activ.(MET hr)	24hr	365	0.69	2.01	4.13	0.256	0.328
Inte. Activ.(MET hr)	24hr	365	1.83	7.16	0.01	0.203	0.999
% SFA(%calories)	24hr	365	6.63	1.34	11.93	0.832	0.101
Total FAT(gm)	24hr	365	695.95	138.14	946.99	0.834	0.127
Total SFA(gm)	24hr	365	113.24	18.56	151.01	0.859	0.109
BMI (kg/m2)	quarter	365	30.06	0.47	0.01	0.984	0.984
DBP (mm Hg)	quarter	365	64.77	36.14	6.19	0.642	0.854
HDL (mg/dl)	quarter	365	127.97	13.30	13.55	0.906	0.495
LDL (mg/dl)	quarter	365	1135.58	44.58	182.30	0.962	0.196
SBP (mm Hg)	quarter	365	231.14	70.78	13.40	0.766	0.841
TC (mg/dl)	quarter	365	1463.33	61.32	225.00	0.960	0.214
TG (mg/dl)	quarter	365	14606.16	3354.32	873.32	0.813	0.793
Weight (kg)	quarter	365	298.35	3.81	0.02	0.987	0.995

*24hr: Based on a 24 hours recall telephone interview given for up to 3 days in a quarter
*7ddr: 7-day dietary records based on subject recall, 1 each per quarter.
Source: tpy04p20j.sas, tpy04p019h.sas, tpy04p23h.sas, tpy04p027l4.sas, tpy04p028i.sas

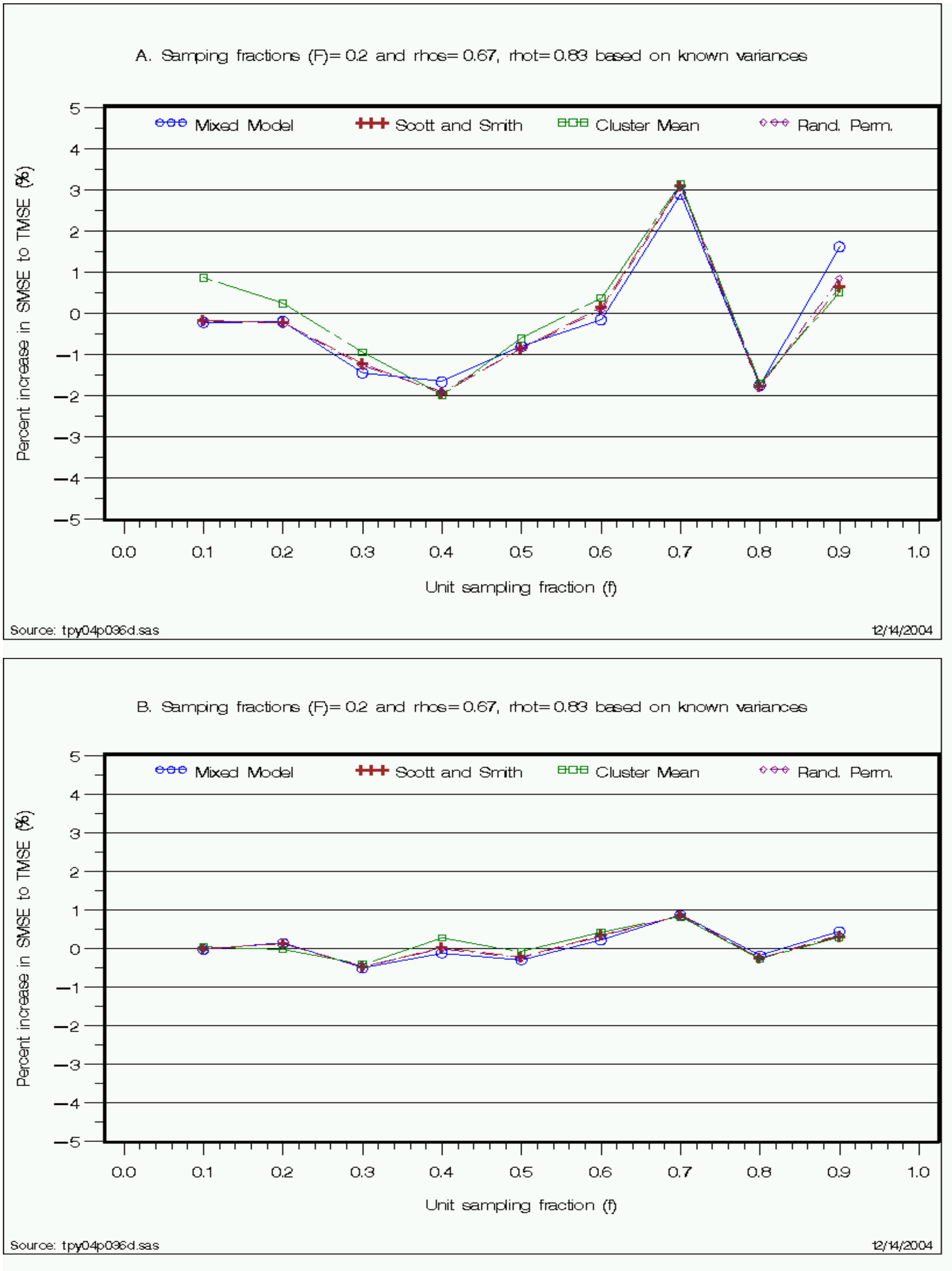
Table 6. Cluster and unit intra-class correlations of some health variables with a time period of 365 days

	unit intra-class correlation (r_t)						
cluster intra-class correlation (r_s)		[0-0.10)	[0.10-0.35)	[0.35-0.65)	[0.65-0.87)	[0.87-1)	1
	[0-0.10)						
	[0.10-0.35)		Moderate Act.			Intense Act.	
	[0.35-0.65)	%FAT**			Total FAT* DBP		
	[0.65-0.87)	Total Fat**	Total SFA** %SFA**	%SFA*, %FAT*, Total SFA*	SBP TG		
	[0.87-1)		LDL, TC	HDL		BMI, WT	
	1	Light Act.					

*:7ddr, 7 days dietary records based on subject recall, 1 each per quarter

**:24hr, based on a 24 hour recall telephone interview given for up to 3 days in a quarter

Figure 1. Increment in SMSE relative to TMSE of Cluster Mean Model, Mixed Model, Scott and Smith, and Random Permutation at two simulation runs (top=1000, bottom=10000) with ρ_s = 0.67, ρ_t = 0.83

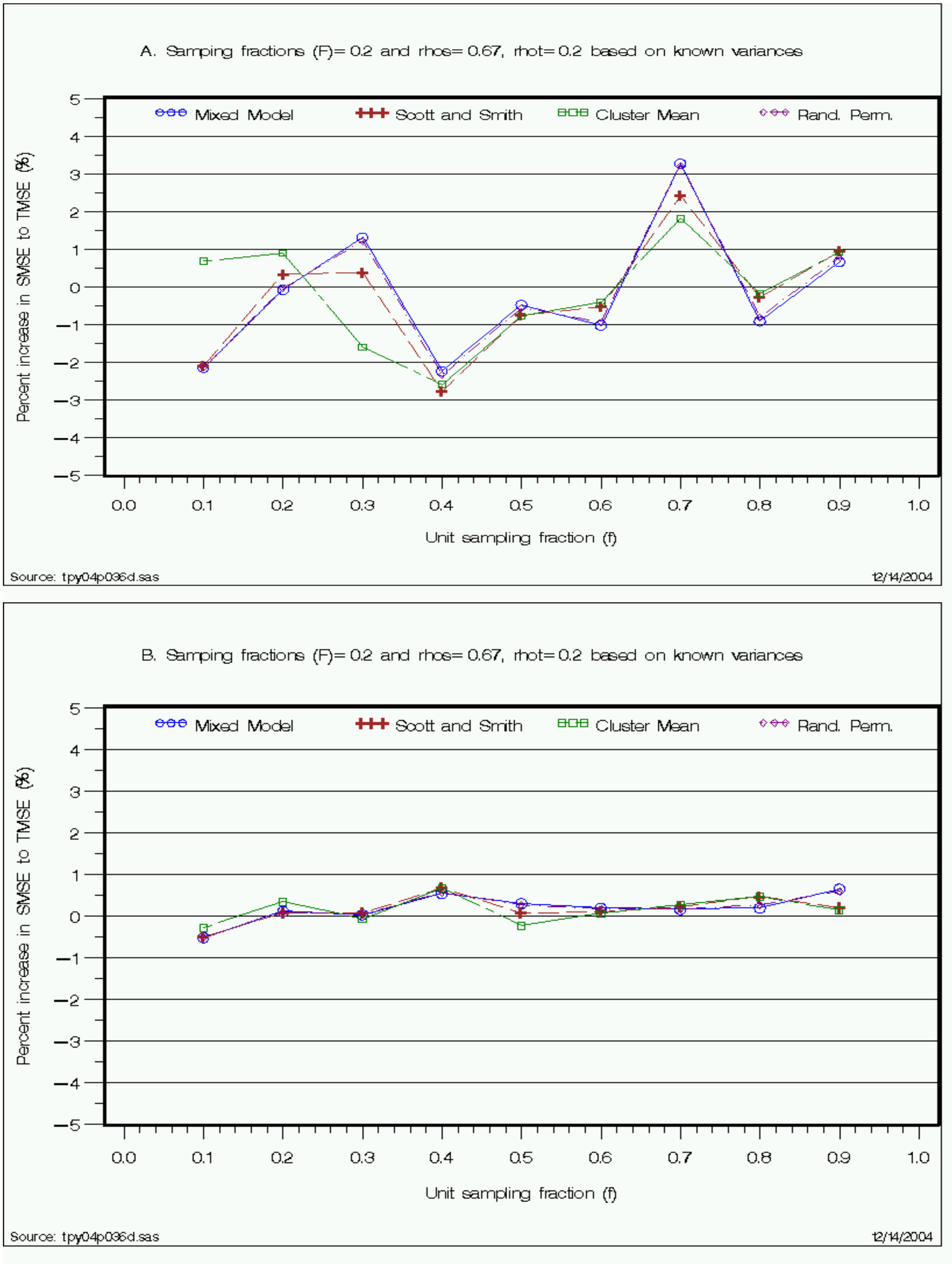

A. Samping fractions (F)= 0.2 and rhos= 0.67, rhot= 0.83 based on known variances

Source: tpy04p036d.sas 12/14/2004

B. Samping fractions (F)= 0.2 and rhos= 0.67, rhot= 0.83 based on known variances

Source: tpy04p036d.sas 12/14/2004

Figure 2. Increment in SMSE relative to TMSE of Cluster Mean Model, Mixed Model, Scott and Smith, and Random Permutation at two simulation runs (top=1000, bottom=10000) with ρ_s = 0.67, ρ_t = 0.20


A. Samping fractions (F)= 0.2 and rhos= 0.67, rhot= 0.2 based on known variances

Source: tpy04p036d.sas 12/14/2004

B. Samping fractions (F)= 0.2 and rhos= 0.67, rhot= 0.2 based on known variances

Source: tpy04p036d.sas 12/14/2004

Figure 3. Increment in TMSE of predictors over TMSE of Random Permutation at one simulation (number of runs=10000) at two cases (top: $\rho_s = 0.67$, $\rho_t = 0.83$, bottom: $\rho_s = 0.67$, $\rho_t = 0.20$) based on known variances

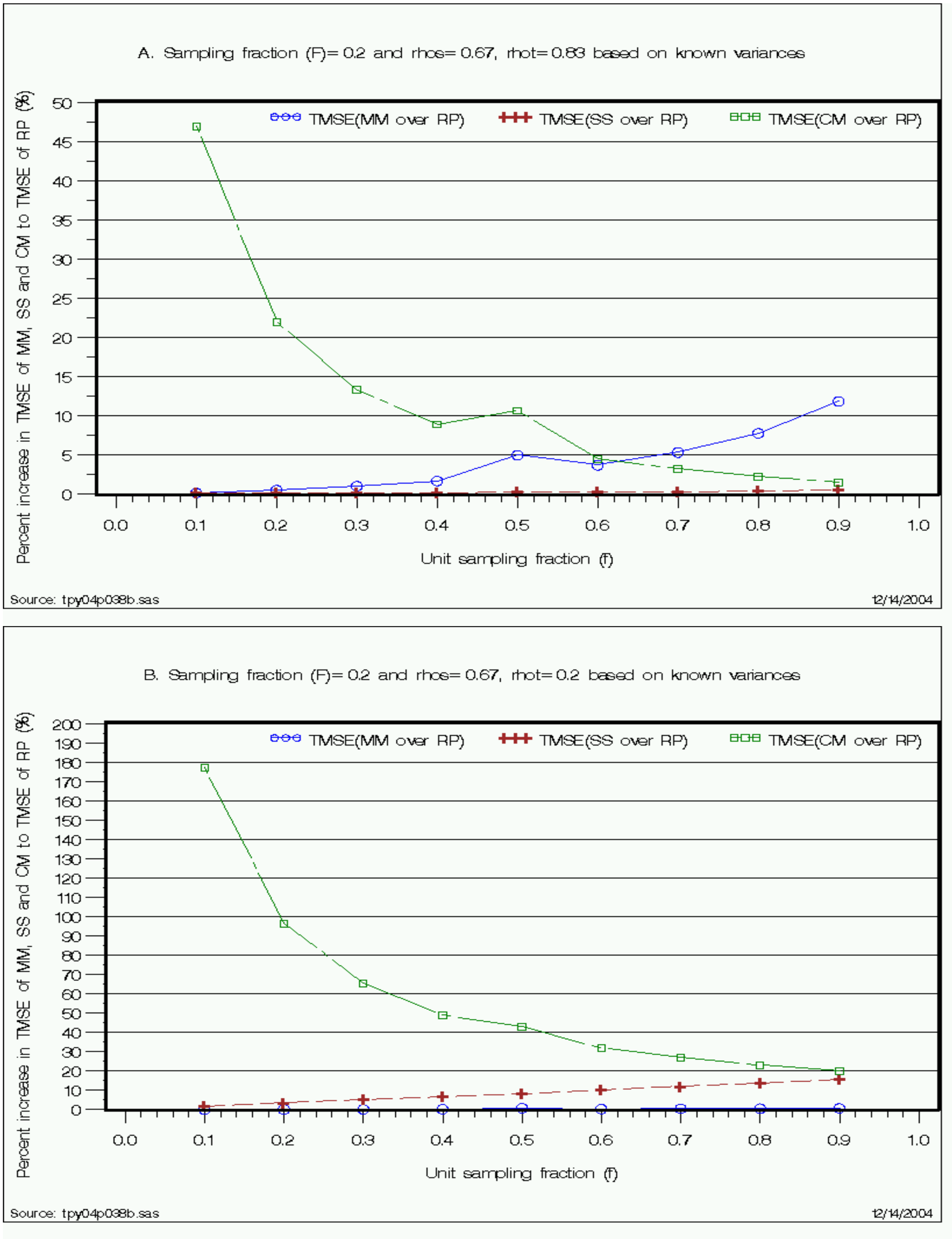
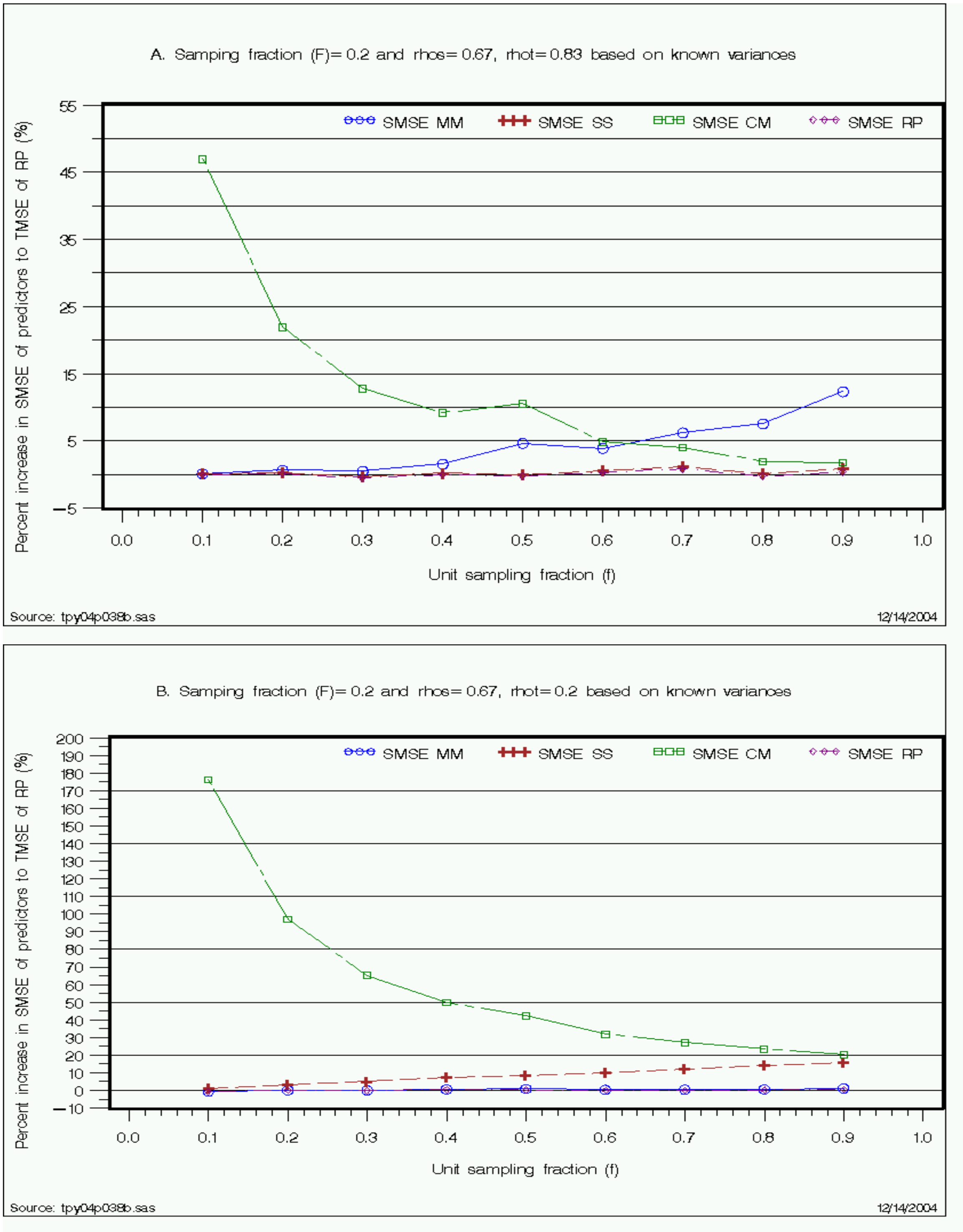


Figure 4. Increment in SMSE of predictors over TMSE of Random Permutation at one simulation (number of runs =10000) at two cases (top: ρ_s = 0.67, ρ_t = 0.83 , bottom: ρ_s = 0.67, ρ_t = 0.20) based on known variances



A. Samping fraction (F)= 0.2 and rhos= 0.67, rhot= 0.83 based on known variances

Source: tpy04p038b.sas

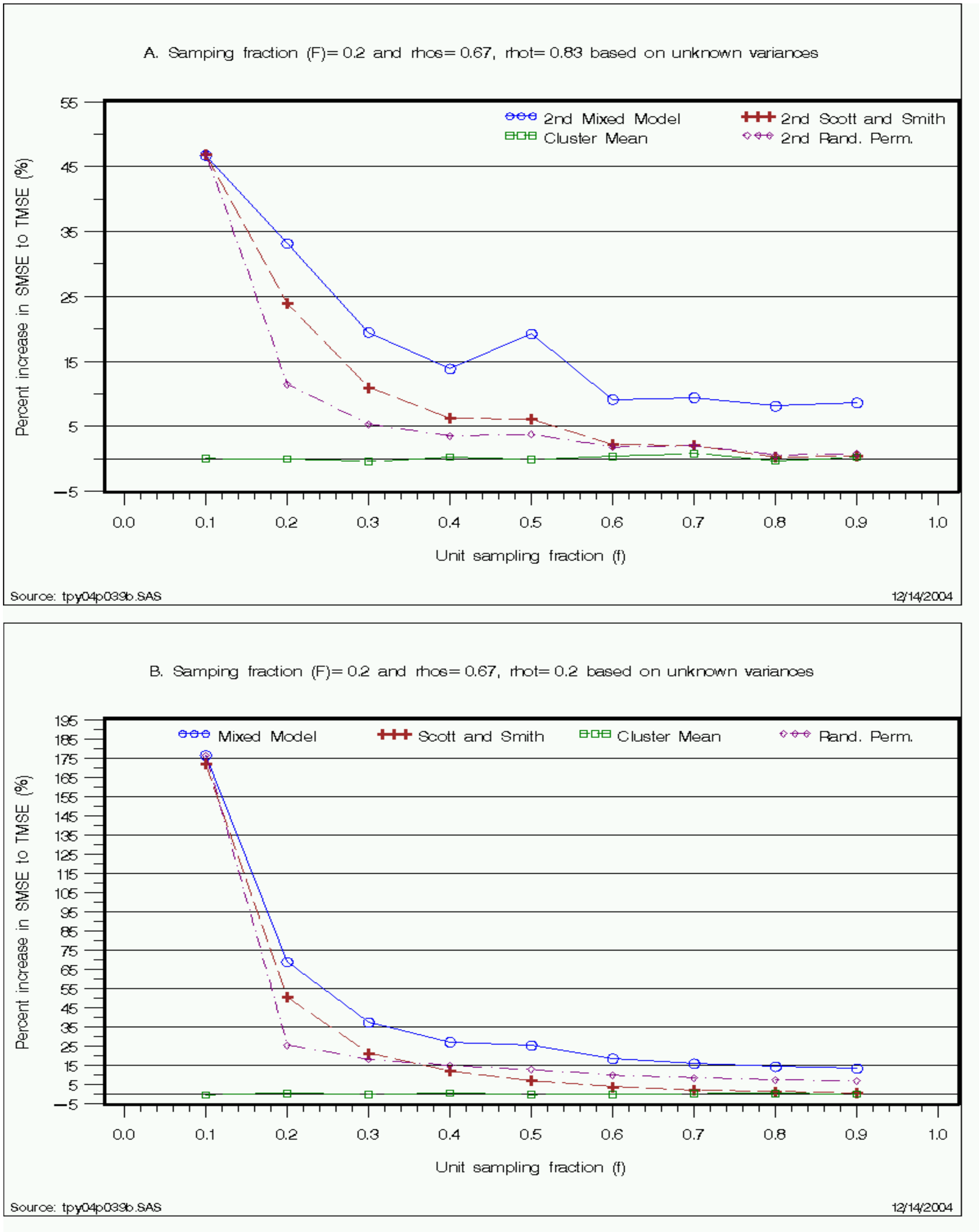
12/14/2004

B. Samping fraction (F)= 0.2 and rhos= 0.67, rhot= 0.2 based on known variances

Source: tpy04p038b.sas

12/14/2004

Figure 5. Increment in SMSE relative to TMSE of Cluster Mean Model, Mixed Model, Scott and Smith, and Random Permutation at one simulation (number of runs=10000) at two cases (top: $\rho_s = 0.67$, $\rho_t = 0.83$, bottom: $\rho_s = 0.67$, $\rho_t = 0.20$) when variances are unknown



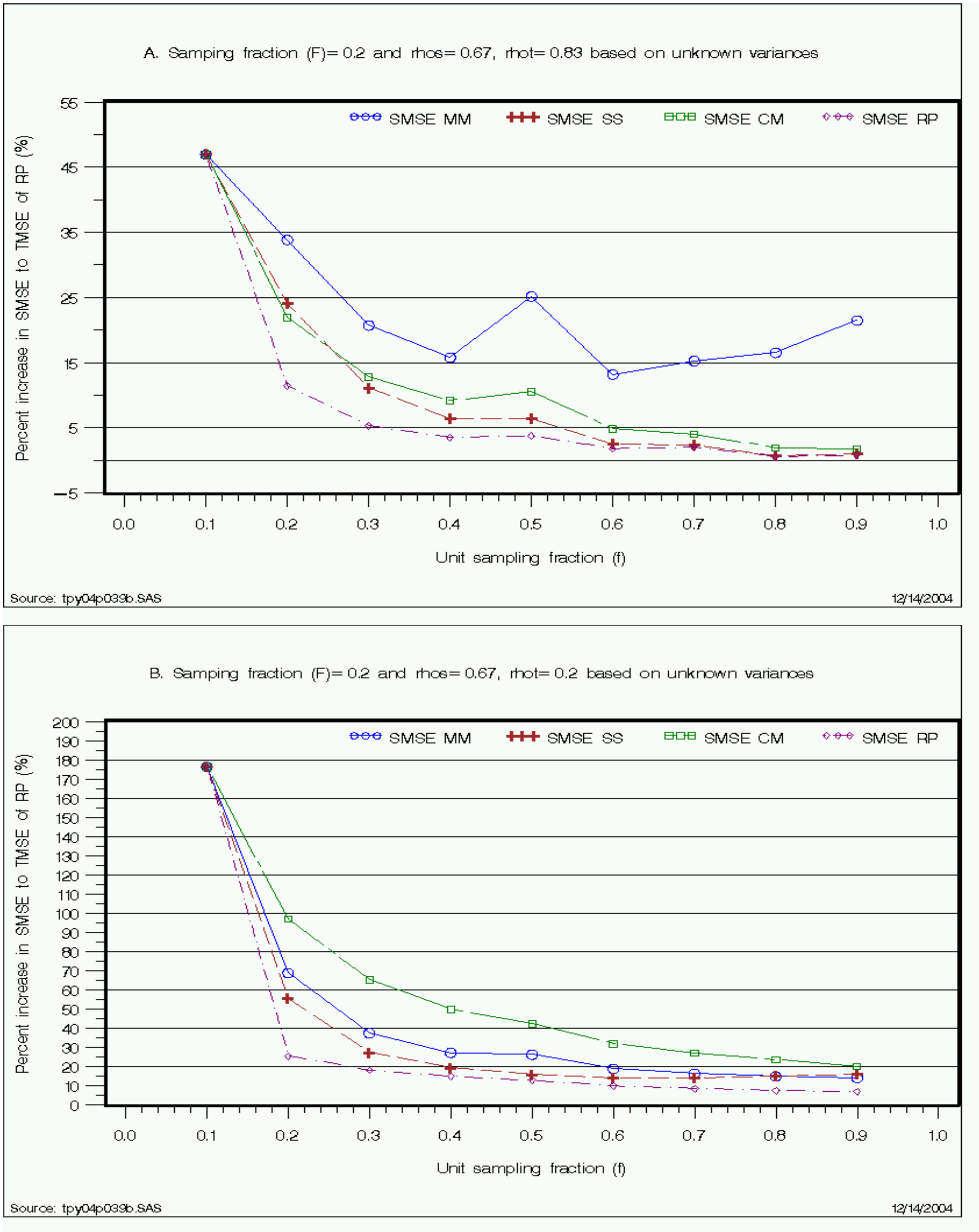
A. Samping fraction (F)= 0.2 and rhos= 0.67, rhot= 0.83 based on unknown variances

Source: tpy04p039b.SAS 12/14/2004

B. Samping fraction (F)= 0.2 and rhos= 0.67, rhot= 0.2 based on unknown variances

Source: tpy04p039b.SAS 12/14/2004

Figure 6. Increment in SMSE of predictors over TMSE of Random Permutation at one simulation (number of runs=10000) at two cases (top: $\rho_s = 0.67$, $\rho_t = 0.83$, bottom: $\rho_s = 0.67$, $\rho_t = 0.20$) when variances are unknown



A. Samping fraction (F)= 0.2 and rhos= 0.67, rhot= 0.83 based on unknown variances

Source: tpy04p039b.SAS 12/14/2004

B. Samping fraction (F)= 0.2 and rhos= 0.67, rhot= 0.2 based on unknown variances

Source: tpy04p039b.SAS 12/14/2004

5. CONCLUSIONS

When variances were known, and the within-cluster variances were equal, there was not much of a difference in the percent increases in SMSE relative to TMSE for four predictors at both settings of cluster and unit intra-class correlations ($r_s = 0.67$, $r_t = 0.83$; r_s=0.67, r_t=0.2). However, when the variances were unknown, and within-cluster variances were equal, there were differences in SMSE relative to TMSE for four predictors in both cases of cluster and unit intra-class correlations, especially at smaller unit sampling fraction (r_t). When variances were known, within-cluster variances were equal, and both cluster and unit intra-class correlations were bigger than 0.5, the Random Permutation Predictor had the minimum TMSE compared to TMSE of other predictors, followed by Scott and Smith's Predictor. When the unit intra-class correlation was less than 0.5, for example, 0.2, and other setting were the same, the Random Permutation Predictor still had the minimum TMSE compared to TMSE of other predictors, but was followed by the Mixed Model. This was consistent with the results given by Stanek and Singer (2004), in which the Random Permutation Predictor had the smallest TMSE. When variances were known, and the within-cluster variances were equal, and both cluster and unit intra-class correlations were greater than 0.5, SMSE of Scott and Smith's Predictor was the closest to the TMSE of the Random Permutation Predictor (Figure 3A). However, when unit intra-class correlation was 0.2, the SMSE of the Mixed Model was the closest to the TMSE of the Random Permutation Predictor (Figure 3B). Martino, Singer and Stanek (2004) reported similar results when variances were known, and the within cluster variances was equal, in which the Random Permutation Predictor had the smallest SMSE, followed by the Mixed Model when ρ_s and ρ_t were small (up to 0.2) or by Scott and Smith's Predictor when ρ_s and ρ_t were moderately large (between 0.5 and 0.8).

When variances were unknown, and within-cluster variances were equal, the Random Permutation Predictor was not always the best among three predictors (excluding the Cluster

Mean) based on the relative MSE $\left(\frac{\text{the simulated MSE-the theoretical MSE}}{\text{the theoretical MSE}}x100\%\right)$ (page 3). Minimum MSE was obtained for the predictor under the Random Permutation when the Cluster sampling fraction (F) and the unit-sampling fraction (f) were small. This was consistent with the results given by Martino, Singer and Stanek (2004). The Cluster Mean Model had the smallest percent increase in SMSE relative to its TMSE among four predictors at both settings of cluster and unit intra-class correlations. Furthermore, The Cluster Mean had a constant relative MSE over the unit-sampling fractions (f). The reason was that the Cluster Mean Model did not depend on the unit-sampling fraction (f) as shown in Table 1.

In addition, when response errors increase, i.e. ρ_t , was small, the relative increase in the predictor's SMSE over the theoretical MSE of the Random Permutation was inflated especially when the unit-sampling fraction was small. That may indicate that performance of predictors became poor at this circumstance.

6. DISCUSSION

There were several limitations to this study. First, data in Season Study were obtained from volunteers, and both subjects and dates of measurements over a quarter were not randomly selected. We assumed that subjects participating in this study to be comparable to a simple random selection from a finite population. In addition, since the response errors of most of the variables in this study were not available in the literature, they were estimated by two simulated responses on each subject. The simulated responses of the variables on each subject depended on both the mean values of variables at quarter 1 of the quarterly data set in the Seasons study and the coefficient of variations of the variables. We had made up coefficients of variation corresponding to simple plausible assumptions. Moreover, we only assumed that data were normally distributed. In addition, we assumed that number of units in each cluster was equal, and an equal number of units were selected from each selected cluster. And we had not simulated results when the unit-sampling fraction (f) was under 0.1. Despite those limitations, this study had several features worthy of our research efforts. This study may be used to assist public health researchers in evaluating BLUP with variables of public health importance in different practical settings. Therefore, this research had provided practical guidance as to how to best predict some common attributes such as saturated fat intake. Finally, the predictor could be evaluated by the theoretical and the simulated MSE.

APPENDIX A

SAS CODE SPECIFICATION 1

Data must be transformed as two rows/measures per subject with column values of subject (id), weight.

```
Proc mixed data=quarter method=reml;
        Class id;
        Model weight=/solution;
        Random id;
Run;
```

The residual error in the variance components is the response error of weight.

APPENDIX B

SAS CODE SPECIFICATION 2

Statistical analysis for model (3.1) using SAS
Data must be transformed as five rows/measures per subject with column variables identifying the subject (id), total cholesterol (tc).

```
Proc mixed data=quarter method=reml;
        Class id;
        Model tc=/solution;
        Random id;
Run;
```

REFERENCES

Andre J.L., J.C. Petit, R. Gueguen, and J.P. Deschamps. 1987. Variability of arterial pressure and heart rate measured at two periods of 15 minutes to 15 days intervals. Arch Mal Coeur Vaiss. June:80(6):1005-1010.

Cavelaars M., J.H. Tulen. J.H. Van Bemmel, P.G. Mulder. And A.H. Van Den Meiracker. 2004. Repoducibility of intra-arterial ambulatory blood pressure: effect of physical activity and posture. Journal of Hypertens. June:22(6)1105-1112.

Goldberger, A. S. 1962. Best linear Unbiased Prediction in the Generalized Linear Regression Model. Journal of the American Statistical Association. 57: 369-375.

Hegested, D.M. and R.J. Nicolosi. 1987. Individual variation in serum cholesterol levels. Proc Natl Acad Sci., USA. 84(17) 6259-6261.

Martino, S. S., J. Singer, E. J. Stanek. 2004. Performance of balanced two-stage empirical predictors of realized cluster latent values from finite population: A simulation study. unpublished report.

Matthews, C.E., J.R. Hebert, P.S. Freedson, E. J. Stanek III, P.A. Merriam, C.B. Ebbeling, and Ira S. Ockene. 2001. Sources of variance in daily physical activity levels in the seasonal variation of blood cholesterol study. American Journal of Epidemiology. Volumn 153. Number 10. 987-995.

Ripolles, Oriti M, Rioboo E. Martin., Moreno A. Diaz., Baena B. Aranguren., Simon M. Murcia., Medina A. Medina., and Del Pozo FJ Fonseca. 2001. Agreement in the measurement of blood pressure among different health professionals. Are mercury sphygmomanometers reliable? Aten Primaria. March 15;27(4)234-243.

Robinson, G. K. 1991. That BLUP is a good thing: the estimation of random effects. Statistical Science 6(1): 15-51.

Royall, R. M. 1976. The linear least squares prediction approach to two-stage sampling. Journal of the American Statistical Association 71:657-664.

Samaniego, F. 2003. Comments to “Predicting Random effects from finite population Clustered samples with response errors authored by Stanek and Singer”.

Sarndal, Carl-Erik, Bengt Swensson and Jan. Wretman. 1991. Model assisted survey Sampling. Springer-Verlag. p124-150.

SAS Institute Inc. 1999. SAS/STAT User’s Guide. Version 8.0. Gary, NC.

Searle, S.R., G. Casella and C.E. McCulloch.1992. Variance components. Wiley Series in Probability and mathematical statistics. John Wiley & Sons, Inc.

Scott, A. and T.M.F. Smith. 1969. Estimation in multi-stage surveys. Journal of the

American Statistical Association 64(327): 830-840.

Smith, S.J., G.R. Cooper, G.L. Myers, and E.J. Sampson. 1993. Biological variability in concentrations of serum lipids: sources of variation among results from published studies and composite predicted values. Clin Chem. June; 39(6)1012-1022.

Stanek, E. J. III., A. Well. And I. Ockene. 1999. Why not routinely use best linear unbiased predictors (BLUPS) as estimates of cholesterol, percent fat from kea and physical activity? Statistics in Medicine 18: 2943-2959.

Stanek, E. J. III, 2003. Evaluating the MSE of Predictors in Balanced Two Stage Predictors of Realized Random Cluster Means With Response Error. (http://www.umass.edu/cluster/ed/biblio-papers.html).

Stanek, E. J. III., Julio Singer. 2004. Predicting random effects from finite population clustered samples with response errors. Journal of the American Statistical Association. December 2004, Vol. 99, No. 468. Theory and Methods.

Stanek, E. 2003a. Notation used to Construct Predictors and Estimates of Predictors in the Simulation Study for Performance of Balanced Two Stage Predictors of Realized Random Cluster Means. (http://www.umass.edu/cluster/ed/biblio-papers.html).

Stanek, E. 2003b. Estimating the Variance in a Simulation Study of Balanced Two Stage Predictors of Realized Random Cluster Means. (http://www.umass.edu/cluster/ed/biblio-papers.html).

Stanek, E. 2003c. Predicting random effects in group randomized trials. (http://www.umass.edu/cluster/ed/biblio-papers.html).