Published in final edited form as:

Statistics (Ber). 2012 January 1; 46(5): 663–671. doi:10.1080/02331888.2010.545408.

Design-based random permutation models with auxiliary information[¶]

WENJUN LI^{†,*}, EDWARD J. STANEK III[‡], and JULIO M. SINGER[§]

[†] University of Massachusetts Medical School, Worcester, MA, USA

[‡] University of Massachusetts, Amherst, MA, USA

[§] Universidade de São Paulo, São Paulo, Brazil

Abstract

We extend the random permutation model to obtain the best linear unbiased estimator of a finite population mean accounting for auxiliary variables under simple random sampling without replacement (SRS) or stratified SRS. The proposed method provides a systematic design-based justification for well-known results involving common estimators derived under minimal assumptions that do not require specification of a functional relationship between the response and the auxiliary variables.

Keywords

auxiliary variable; design-based inference; prediction; finite sampling; random permutation model; simultaneous permutation

1. Introduction

Improvement in the precision of estimates of population parameters based on random samples can be made, at least theoretically, by accounting for auxiliary information (e.g., age, gender etc.). The amount of improvement depends on the correlation between the response and the auxiliary variables. Since the sample means of the auxiliary variables are not likely to equal their population means, a factor proportional to the extent of this difference can be used to adjust the response sample mean.

Although adjusting for auxiliary information has a broad appeal, justification of the adjusted estimator for a finite population typically requires various assumptions beyond sampling [1]. For example, to construct optimal estimators in a model-assisted approach [2], a linear regression model relating the response to the auxiliary variables is assumed for a superpopulation. In this context, optimality is determined by minimizing the expected value of the design-based mean squared errors (MSE) over the superpopulation model. This approach leads to the well-known generalized regression (GREG) estimators [3].

In a second approach, known as calibration, weighted linear estimators that have minimal variance are developed subject to the constraint that the weighted sample auxiliary variable total matches its population counterpart [4].

*Correspondence: Wenjun Li, PhD, (Wenjun.Li@UMASSMED.EDU), Division of Preventive and Behavioral Medicine, University of Massachusetts Medical School, Shaw Building SH2-230, 55 Lake Avenue North, Worcester, MA 01655.

Alternatively, in the so-called prediction approach, predictors of the mean are based on a model (either linear or ratio) relating the response to the auxiliary variables, and on the joint distribution of the sample and the remaining random variables[5]. The predictor is based only on model assumptions, and the sampling scheme plays no role in its derivation. The major disadvantages of this approach include the need for assumptions that cannot be confirmed, and the lack of connection between the model and the physical problem under investigation.

Other approaches are also reported in literature. For example, the empirical likelihood based method proposed by Hartley and Rao [6] is essentially motivated by model assumptions linking the response to the auxiliary variables in the sample, or by a limiting large sample approximation to the hypergeometric model. Using large sample theory arguments, additional estimators that adjust for the conditional bias have been proposed [7]. Other authors, as for example, Holt and Smith [8], considered estimators based on conditional criteria, arguing that a conditional framework is more appropriate than an unconditional one.

Although many of the estimators obtained under these different approaches are identical, their development either requires assumptions beyond those pertaining to the sample design or lack an integrated theory. We develop a design-based estimator of a linear function of the response that accounts for auxiliary information, and requires no assumptions beyond those considered in simple random sampling. The development extends the random permutation model proposed by Stanek et al. [9, 10] to account for auxiliary variables. Under this minimal assumption setup, we show that the optimal estimator is identical to the estimator given by Cochran [11], but does not require the accompanying linear model assumptions. The best linear unbiased estimator (BLUE) is the sum of two terms, one corresponding to the contribution from the sample, and the other corresponding to a best linear unbiased predictor (BLUP) of the contribution due the remaining unobserved random variables.

We begin with the definitions and notation for the random permutation model, and add multiple auxiliary variables. We then derive the BLUE of the population mean under SRS, and subsequently extend the results to stratified SRS. We illustrate the practical application of the method in a study on nutrition among hospital workers in the United States. Finally, we highlight the flexibility, advantages and limitations of the design-based prediction model framework, and comment on directions for future research.

2. A Design Based Model for Simple Random Sampling

Briefly, the model proposed by Stanek et al. [9] uses indicator random variables to represent the permutation of the units in the population. The stochastic model that underlies these random variables is hence referred to as a random permutation model [9] but is different from those considered in Cochran [11] Kempthorne [12] or from those by Rao and Bellhouse [13], Rao [14], and Rao [15].

We illustrate how the random permutation proposed by Stanek et al. [9] can be extended to include multivariate auxiliary information under SRS and stratified SRS. In this context, we develop an optimal estimator of a parameter defined in the population, assuming that the population means of the auxiliary variables are known.

Let the population consist of N subjects, indexed by $s = 1, 2, \dots, N$ non-informative labels. Associated with subject s is a non-stochastic potentially observable $(Q + 1) \times 1$ vector,

$\mathbf{z}_s = \begin{pmatrix} y_s & \mathbf{x}_s' \end{pmatrix}'$, where y_s denotes the outcome of interest and the $Q \times 1$ vector $\mathbf{x}_s$ contains the values of the auxiliary variables. We represent the $(Q + 1) \times 1$ vector of population means

by $\mu = (\mu_y \quad \mu'_x)'$, where $\mu_y = \frac{1}{N} \sum_{s=1}^N y_s$ and $\mu'_x = \frac{1}{N} \sum_{s=1}^N \mathbf{x}'_s$, and the corresponding $(Q+1) \times (Q+1)$ covariance matrix, by $\Sigma = \frac{1}{N-1} \sum_{s=1}^N (\mathbf{z}_s - \mu)(\mathbf{z}_s - \mu)' = \begin{pmatrix} \sigma_y^2 & \sigma'_{yx} \\ \sigma_{yx} & \Sigma_x \end{pmatrix}$, where σ_{yx} is the $Q \times 1$ vector of covariances between the response and the auxiliary variables and Σ_x is the $Q \times Q$ covariance matrix of the auxiliary variables.

We define the random permutation model as the set of all possible equally likely permutations of subjects in the population and represent a random permutation by the sequence of $Q \times 1$ random vectors, $\mathbf{Z}_i$, $i = 1, \dots, N$ where the subscript refers to the position in the permutation. Following Stanek et al. [9], we explicitly construct these random variables using a set of indicator random variables U_{is} , $i = 1, 2, \dots, N$, that have a value of one if subject s is in position i in a permutation, and zero otherwise. Using this notation,

$\mathbf{Z}_i = \sum_{s=1}^N U_{is} \mathbf{z}_s = \mathbf{z}' \mathbf{U}_i$, where $\mathbf{U}_i = (U_{i1} \quad U_{i2} \quad \dots \quad U_{iN})'$ is an $N \times 1$ vector and $\mathbf{z} = ((\mathbf{z}'_s))$ is an $N \times (Q+1)$ matrix, so that the random permutation is represented by the $N \times (Q+1)$ matrix

$\mathbf{Z} = ((\mathbf{Z}_i)) = \mathbf{U} \mathbf{z}$ where $\mathbf{U} = (\mathbf{U}_1 \quad \mathbf{U}_2 \quad \dots \quad \mathbf{U}_N)$ is an $N \times N$ matrix. We represent

$\mathbf{Z} = \begin{pmatrix} \mathbf{Y} & \mathbf{X} \end{pmatrix}$ where $\mathbf{Y}$ is an $N \times 1$ vector with the values of the response variable, and $\mathbf{X}$ is an $N \times Q$ matrix with the values of the auxiliary variables. We require each realization to be equally likely, subject to the constraint that a single unit is allocated to each position, i.e., $\mathbf{U} \mathbf{1}_N = \mathbf{1}_N$, where $\mathbf{1}_N$ is an $N \times 1$ vector with all elements equal to 1, and all units are

assigned to a position, namely, $\mathbf{1}'_N \mathbf{U} = \mathbf{1}'_N$. Taking the expectation over all possible permutations, we have $E(\mathbf{Z}) = \mathbf{1}_N \mu'$, and the corresponding covariance matrix is $\text{cov}(\text{vec}(\mathbf{Z})) = \Sigma \otimes \mathbf{P}_{N,N}$, where $\otimes$ denotes the Kronecker product, $\mathbf{P}_{a,b} = \mathbf{I}_a - b^{-1} \mathbf{J}_a$, $\mathbf{P}_{a,b} = \mathbf{P}_a$ when $a = b$, $\mathbf{I}_N$ is an $N \times N$ identity matrix, and $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}'_N$. Since the means of the auxiliary variables, μ_{xq} , $q = 1, \dots, Q$ are known for the population, we pre-multiply $\mathbf{X}$ by $\mathbf{P}_N$ to centre the auxiliary variables at zero. It follows that $E \begin{pmatrix} \mathbf{Y} & \mathbf{P}_N \mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_N & \mathbf{0}_{N \times Q} \end{pmatrix} \mu_y$ and $\text{cov} \begin{pmatrix} \mathbf{Y} & \mathbf{P}_N \mathbf{X} \end{pmatrix} = \Sigma \otimes \mathbf{P}_{N,N}$. The sample corresponds to the first n rows of

$\begin{pmatrix} \mathbf{Y} & \mathbf{P}_N \mathbf{X} \end{pmatrix}$, with the remainder corresponding to the subsequent $N - n$ rows. We represent the $n(Q+1)$ sample random variables in the vector $\mathbf{Z}_I = \text{vec} \left[\begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n \times (N-n)} \end{pmatrix} \begin{pmatrix} \mathbf{Y} & \mathbf{P}_N \mathbf{X} \end{pmatrix} \right]$, and the $(N-n)(Q+1)$ remainder random variables in the vector

$\mathbf{Z}_{II} = \text{vec} \left[\begin{pmatrix} \mathbf{0}_{(N-n) \times n} & \mathbf{I}_{N-n} \end{pmatrix} \begin{pmatrix} \mathbf{Y} & \mathbf{P}_N \mathbf{X} \end{pmatrix} \right]$. This notation explicitly represents the SRS process by a stochastic model. The expected value of the sample and remainder random variables are

$E(\mathbf{Z}_I) = \mathbf{G}_I \mu_y$ and $E(\mathbf{Z}_{II}) = \mathbf{G}_{II} \mu_y$, where $\mathbf{G}_I = \begin{pmatrix} \mathbf{1}'_n & \mathbf{0}'_{nQ} \end{pmatrix}$ and $\mathbf{G}_{II} = \begin{pmatrix} \mathbf{1}'_{N-n} & \mathbf{0}'_{(N-n)Q} \end{pmatrix}$. The

corresponding covariance matrix is $\text{var} \begin{pmatrix} \mathbf{Z}_I \\ \mathbf{Z}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_I & \mathbf{V}_{I,II} \\ \mathbf{V}'_{I,II} & \mathbf{V}_{II} \end{pmatrix}$, where $\mathbf{V}_I = \Sigma \otimes \mathbf{P}_{n,N}$ has dimension $n(Q+1) \times n(Q+1)$, $\mathbf{V}_{II} = \Sigma \otimes \mathbf{P}_{(N-n),N}$ has dimension $(N-n)(Q+1) \times (N-n)(Q+1)$, and $\mathbf{V}_{I,II} = -\frac{1}{N} \sum \otimes \mathbf{J}_{n \times (N-n)}$ has dimension $n(Q+1) \times (N-n)(Q+1)$, with

$\mathbf{J}_{n \times (N-n)} = \mathbf{1}_n \mathbf{1}'_{N-n}$. Consequently, the partitioned model that reflects SRS can be represented as

$$\begin{pmatrix} \mathbf{Z}_I \\ \mathbf{Z}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{G}_I \\ \mathbf{G}_{II} \end{pmatrix} \mu_y + \mathbf{E}, \quad (1)$$

where $\mathbf{E}$ is the vector of residuals.

3. BLUE of a linear function of the parameters

Our interest lies in linear combinations of the parameters which may also be expressed as linear combinations of the permuted response variable, namely,

$$\tau = \mathbf{c}'\mathbf{Y} = \sum_{i=1}^N c_i Y_i = \mathbf{C}'_I \mathbf{Z}_I + \mathbf{C}'_{II} \mathbf{Z}_{II} \quad (2)$$

where $\mathbf{C}_I = \begin{pmatrix} \mathbf{c}'_I & \mathbf{0}'_{nQ \times 1} \end{pmatrix}$, $\mathbf{C}_{II} = \begin{pmatrix} \mathbf{c}'_{II} & \mathbf{0}'_{(N-n)Q \times 1} \end{pmatrix}$ and $\mathbf{c}_i = \begin{pmatrix} c'_i & c'_{II} \end{pmatrix}$. When $c_i = \frac{1}{N}$, $i = 1, \dots, N$, it follows that $\tau = \mu_Y$ (the population mean); when $c_i = 1$, $i = 1, \dots, N$, τ is the population total. Since only $\mathbf{C}'_{II} \mathbf{Z}_{II}$ will be unknown after sampling, estimating τ is equivalent to predicting this term. Following Royall's prediction approach [16], we develop the BLUP of $\mathbf{C}'_{II} \mathbf{Z}_{II}$ which, when added to $\mathbf{C}'_I \mathbf{Z}_I$, is the BLUE of τ .

We require the predictor to i) be a linear function of the sample, i.e., to be of the form $\mathbf{w}'\mathbf{Z}_I$, where $\mathbf{w}$ is an $n(Q+1) \times 1$ vector of unknown coefficients, so that the estimator of τ is of the form $T_y = \mathbf{C}'_I \mathbf{Z}_I + \mathbf{w}'\mathbf{Z}_I$; ii) be an unbiased predictor of $\mathbf{C}'_{II} \mathbf{Z}_{II}$, i.e., $E(\mathbf{w}'\mathbf{Z}_I) = E(\mathbf{C}'_{II} \mathbf{Z}_{II})$; and iii) have minimum expected MSE. Applying Royall's prediction theorem [16], we obtain

$$\widehat{\mathbf{w}} = N\bar{c}_{II} \left((1-f) \frac{\mathbf{1}'_n}{n} \mathbf{I} - \boldsymbol{\beta}' \otimes \frac{\mathbf{1}'_n}{n} \right) \quad (3)$$

where $\bar{c}_{II} = \frac{1}{N-n} \sum_{i=n+1}^N c_i$, $f = n/N$ and $\boldsymbol{\beta} = \sum_X^{-1} \boldsymbol{\sigma}_{yX} = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_Q \end{pmatrix}'$. Details of the derivation are given in Appendix I. Consequently,

$$\widehat{T}_y = \sum_{i=1}^n c_i Y_i + (N-n)\bar{c}_{II} \left[\bar{Y}_I - \frac{1}{1-f} \sum_{q=1}^Q \beta_q (\bar{X}_{qI} - \mu_q) \right] \quad (4)$$

where $\bar{Y}_I$ and $\bar{X}_{qI}$, $q = 1, \dots, Q$ are sample means. The first component of (4) corresponds to the contribution from the observed sample, and the second is the BLUP of the remainder part.

The variance is given by

$$\text{var}(\widehat{T}_y) = \frac{N^2}{n} \sigma_y^2 (1 - \rho_{yX}^2) (1-f)\bar{c}_{II}^2 + (N-n-1)\sigma_y^2 \sigma_{c_{II}}^2, \quad (5)$$

where $\sigma_{c_{II}}^2 = (N-n-1)^{-1} \sum_{i=n+1}^N (c_i - \bar{c}_{II})^2$, and $\rho_{yX}^2 = \boldsymbol{\sigma}'_{yX} \sum_X^{-1} \boldsymbol{\sigma}_{Xy} / \sigma_y^2$ is the squared multiple correlation coefficient of response and auxiliary variables. In practical applications, $\boldsymbol{\beta}$, σ_y^2 and ρ_{yX}^2 are not known, and must be replaced by estimates. While $\boldsymbol{\sigma}_{Xy}$ is usually unknown in practice, $\boldsymbol{\Sigma}_X$ may be known. Use of $\boldsymbol{\Sigma}_X$ rather than $\widehat{\boldsymbol{\Sigma}}_X$ will avoid the potential problem of

having a singular covariance matrix estimated from the sample and result in estimators with smaller variance, especially when the sample size is small [17].

4. Extensions to stratified simple random sampling

We use notation similar to that in Section 2, and add the index $h = 1, 2, \dots, H$ to extend the results to stratified SRS. Let N_h and n_h be the h -th stratum population and sample sizes, respectively, $f_h = n_h/N_h$ be the sampling fraction, y_{hs} be the value of the response for subject

s in stratum h , $\mu_{hy} = N_h^{-1} \sum_{s=1}^{N_h} y_{hs}$ be the stratum mean of the response, and $\mu_y = \frac{1}{N} \sum_{h=1}^H \sum_{s=1}^{N_h} y_{hs}$ be the corresponding population mean. Also let x_{hqs} denote the value of the q -th auxiliary variable for subject s in stratum h , $\bar{x}_{hq} = n_h^{-1} \sum_{s=1}^{n_h} x_{hqs}$ be the corresponding stratum sample mean, $\boldsymbol{\mu}_h = \begin{pmatrix} \mu_{hy} & \mathbf{0}_{1 \times Q} \end{pmatrix}'$ be the vector of stratum means for all $Q+1$ variables (after subtracting the stratum auxiliary variable means, assumed to be known) in stratum h , and $\boldsymbol{\Sigma}_h$ be the corresponding stratum covariance matrix.

We represent the joint permutation of the response and the Q auxiliary variables in stratum h , $h = 1, \dots, H$, as an $N_h \times (Q+1)$ matrix $\mathbf{Z}_h = \mathbf{U}_h \mathbf{z}_h$, where $\mathbf{z}_h$ is an $N_h \times (Q+1)$ matrix and $\mathbf{U}_h$ is the corresponding $N_h \times N_h$ matrix of random indicator variables. By definition, sampling is independent between strata, so that $\mathbf{U}_h$ is independent from $\mathbf{U}_{h^*}$ for $h \neq h^*$. As a result, $E(\mathbf{Z}_h) = \mathbf{1}_{N_h} \boldsymbol{\mu}_h'$, and $\text{var}[\text{vec}(\mathbf{Z}_h)] = \boldsymbol{\Sigma}_h \otimes \mathbf{P}_{N_h}$; additionally, all covariances between strata are zero. Suppose that a sample of size n_h is drawn by SRS from stratum h , $h = 1, \dots, H$. Model (1) can be extended to represent stratified SRS by defining the

$\left(\sum_{h=1}^H n_h(Q+1) \right) \times 1$ sample random variables as $\mathbf{Z}_I = \begin{pmatrix} \mathbf{Z}'_{I1} & \mathbf{Z}'_{I2} & \dots & \mathbf{Z}'_{IH} \end{pmatrix}'$, and the remaining $\left(\sum_{h=1}^H (N_h - n_h)(Q+1) \right) \times 1$ random variables as $\mathbf{Z}_{II} = \begin{pmatrix} \mathbf{Z}'_{II1} & \mathbf{Z}'_{II2} & \dots & \mathbf{Z}'_{IIH} \end{pmatrix}'$. Terms in these vectors are defined by the $n_h(Q+1)$ sample random variables,

$\mathbf{Z}_{Ih} = \text{vec} \left[\begin{pmatrix} \mathbf{I}_{n_h} & \mathbf{0}_{n_h \times (N_h - n_h)} \end{pmatrix} \mathbf{Z}_h \right]$ and the $(N_h - n_h)(Q+1)$ remainder random variables $\mathbf{Z}_{IIh} = \text{vec} \left[\begin{pmatrix} \mathbf{0}_{(N_h - n_h) \times n_h} & \mathbf{I}_{N_h - n_h} \end{pmatrix} \mathbf{Z}_h \right]$ in stratum h , for $h = 1, \dots, H$. The other terms in the

extension of model (1) to stratified SRS are given by $\mathbf{G}_I = \bigoplus_{h=1}^H (\mathbf{I}_{Q+1} \otimes \mathbf{1}_{n_h})$ and

$\mathbf{G}_{II} = \bigoplus_{h=1}^H (\mathbf{I}_{Q+1} \otimes \mathbf{1}_{N_h - n_h})$, where $\bigoplus_{h=1}^H \mathbf{A}_h$ is a block diagonal matrix with diagonal blocks given by $\mathbf{A}_h$ for $h = 1, \dots, H$. The population mean (target parameter) may be expressed as

$$\tau = \mathbf{C}'_I \mathbf{Z}_I + \mathbf{C}'_{II} \mathbf{Z}_{II}, \text{ where } \mathbf{C}_I = \left(\bigoplus_{h=1}^H \left(\begin{pmatrix} 1 & \mathbf{0}_{1 \times Q} \end{pmatrix}' \otimes \frac{\mathbf{1}_{n_h}}{N_h} \right) \right) \mathbf{1}_H \text{ and}$$

$$\mathbf{C}_{II} = \left(\bigoplus_{h=1}^H \left(\begin{pmatrix} 1 & \mathbf{0}_{1 \times Q} \end{pmatrix}' \otimes \frac{\mathbf{1}_{N_h - n_h}}{N_h} \right) \right) \mathbf{1}_H.$$

The BLUE for τ can be derived using the same steps illustrated in Section 3. We consider

linear estimators of the form $\tau = (\mathbf{C}'_I + \mathbf{w}') \mathbf{Z}_I$, where $\mathbf{w}$ is a vector of unknown coefficients.

The unbiasedness constraint for τ requires that $E(\mathbf{w}' \mathbf{Z}_I - \mathbf{C}'_{II} \mathbf{Z}_{II}) = 0$. Following the same optimization procedure described previously, we obtain

$$\widehat{\mathbf{w}} = - \left(\bigoplus_{h=1}^H \left(\begin{pmatrix} 1 \\ \mathbf{0}_{Q \times 1} \end{pmatrix} \otimes \frac{\mathbf{1}_{n_h}}{N_h} \right) \right) \mathbf{1}_H + \left\{ \bigoplus_{h=1}^H \left(\begin{pmatrix} 1 \\ -\boldsymbol{\beta}_h \end{pmatrix} \otimes \frac{\mathbf{1}_{n_h}}{n_h} \right) \right\} \mathbf{1}_H,$$

where $\boldsymbol{\beta}_h = \sum_{hX}^{-1} \boldsymbol{\sigma}_{hXy}$ are the linear coefficients of the regression of the response on the auxiliary variables in the h -th stratum, with $\boldsymbol{\sigma}_{hXy}$ and $\boldsymbol{\Sigma}_{hX}$ each having a similar interpretation as in the SRS case. Consequently, the optimal linear estimator for T_y is

$$\widehat{T}_y = \sum_{h=1}^H \frac{N_h}{N} f_h \bar{y}_h + \sum_{h=1}^H \frac{N_h}{N} (1 - f_h) \left[\bar{y}_h - \frac{1}{1 - f_h} \sum_{q=1}^Q \beta_{hq} (\bar{x}_{hq} - \mu_{hxq}) \right], \quad (6)$$

where $\bar{y}_h$ and $\bar{x}_{hq}$, $q = 1, \dots, Q$, are sample means of the response and auxiliary variables for the h -th stratum. Expression (6) can be further simplified to

$$\widehat{T}_y = \sum_{h=1}^H \frac{N_h}{N} \left[\bar{y}_h - \sum_{q=1}^Q \beta_{hq} (\bar{x}_{hq} - \mu_{hxq}) \right] \quad (7)$$

with the corresponding variance

$$\text{var}(\widehat{T}_y) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 (1 - \rho_h^2) (1 - f_h) n_h^{-1} \sigma_{hy}^2, \quad (8)$$

where $\rho_h^2 = \boldsymbol{\sigma}_{hXy}' \sum_{hX}^{-1} \boldsymbol{\sigma}_{hXy} / \sigma_y^2$ is the squared multiple correlation coefficient between the response variable and auxiliary variables in the h -th stratum.

5. Example

We illustrate the methods using data from the Step Ahead Study, a nutrition and exercise study among hospital workers in a hospital affiliated with the University of Massachusetts Medical School, Worcester, Massachusetts in 2005 [18]. The target is the average fruit and vegetable (FV) consumption of the hospital employees. The consumption is measured by a fruit and vegetable consumption index, ranging from 0 to 28, with 0 representing no consumption and higher values indicating more consumption. The study population consists of 2,553 eligible employees aged between 18 and 65 years, with 79% female, 87% white and 13% racial minorities, and mean age of 43 years. Sex, age, race and job characteristics of all eligible employees are known from hospital administrative data. The sample was drawn using a simple random sampling scheme stratified by sex and minority status (white versus non-white minorities). Men and racial minorities of both genders were purposely oversampled to preserve the ability to estimate means of fruit and vegetable consumption by gender and minority subgroups. Information on the study population and the sampling design is summarized in Table 1.

Investigators were interested in stratum-specific as well as overall mean fruit and vegetable consumption adjusting for known auxiliary information on age. The stratum-specific and overall estimates of consumption indices are presented in Table 2. The BLUE of stratum-specific means of the consumption indices are calculated via (4) and (5). Applying (7) and

(8), the overall mean consumption index estimate and its standard error are given by 11.69 (0.390) in Table 2. The adjusted stratum-specific and overall estimates of the mean consumption index have smaller standard errors than the crude estimates. Larger reductions of the standard errors occurred in the two strata corresponding to minority men and women.

A further analysis suggests that variation in $\hat{\beta}_h$ across strata is substantial. The association of fruit and vegetable consumption with age is stronger among ethnoracial minorities, with the strongest association among minority men. As a result, adjusting for age resulted in substantial variance reduction for stratum-specific estimates for minority men and women, but not for white men and women. Among the four strata, the largest adjustment and the largest variance reduction correspond to the stratum of minority men, where the association between vegetable consumption and age was the strongest and the difference in average age between the sample and the population was the largest (−3.3 years). Because the study population is predominantly white (87%), however, we do not see a significant variance reduction for the estimate of overall population.

5. Discussion

In the context of SRS and stratified SRS with auxiliary information, we obtained BLUE of linear combinations of population parameters using a model with minimal assumptions. Our result is similar to an estimator commonly employed in multiple linear regression models, but includes a finite population correction factor in the variance. In particular, (4) and (5) are identical to difference estimators with optimal coefficients [19], and to the GREG estimator [3]. However, assumptions needed for the derivation require neither a parametric distribution nor the specification of the functional relationship between the response and the auxiliary variables.

The survey sampling literature has struggled to reconcile design-based and model-based theories of estimation/prediction. Model-based methods recently popularized by Valiant et al. [20] stem from the prediction approach developed by Royall [16, 21]. The underlying theoretical structure is important, since it allows such methods to be extended relatively easily to different applications with increasing complexity. A limitation of the theory is that it does not account for the sample design.

A similar unifying theory has not been developed for design-based methods. Cochran's [11] original approach was to postulate a linear regression model, and then determine the regression coefficients based on minimizing the variance. Other approaches, such as the GREG or the calibration approaches [3] combine model-based and design-based ideas or use ad hoc functional forms of estimators, optimizing them in special settings. These approaches have been successful in addressing many practical problems in design-based frameworks [3]. However, they have not provided a consistent conceptual and theoretical basis that can be readily extended to more complex applications.

Representing the sample design via a random permutation model, and then predicting functions of unobserved subjects in a systematic way provides an appealing, straightforward foundation for finite population inference. There are steps in this process that break with tradition, such as expressing a parameter as a sum of random variables. Focusing attention on predicting unobserved quantities is certainly intuitively satisfying, but unusual in the context of estimation.

The estimation methods presented in this paper afford great flexibility for handling various complex practical situations under SRS and stratified SRS. The models can be extended to obtain the BLUE of the population mean when a) multiple response variables are present, b) some, but not all stratum-specific auxiliary means are known, c) availability of auxiliary

information differ between strata, and d) stratum-specific sample sizes are small. Some examples are available at our project website (<http://www.umass.edu/cluster/ed/index.html>).

We have illustrated how the design-based random permutation model theory can be extended to include auxiliary variables in a straightforward manner. The previous developments of the theory have identified subtleties in interpreting random effects in simple random sampling [9] and developed predictors of realized random effects in balanced two stage sampling problems with response error [10]. Current research is directed to extending these results to clustered population settings where clusters are of different size and there is unequal probability sampling, to settings where there are multiple correlated responses variables, and to settings where there are missing data. In each case, a similar approach is considered, with estimators (or predictors) developed via a clear optimization theory.

Our results illustrate how sampling theory, via the random permutation model, can be directly used to account for covariates. In practice, covariance terms in the estimator formulae need to be estimated. Some simulation study results on the impact of such estimation are given by Li [17]. The resulting estimator coincides with those developed by GREG or calibration approaches, and strengthens the appeal of the random permutation model. Still, more work is needed to extend the methods to more complex settings, including auxiliary variables measured with errors, time-dependent auxiliary information in longitudinal studies, and two stage designs with cluster and unit covariates. The results presented here provide the foundation for additional work in these directions.

Acknowledgments

This research was developed with the support of the National Institutes of Health (NIH-PHS-R01-HD36848, R01-HL071828-02 and 5R01HL079483), USA, and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil. We thank Dr. Stephenie Lemon at the University of Massachusetts Medical School for providing example data from the Step Ahead Study (NIH/NHLBI 5R01HL079483).

References

1. Rao JNK. Developments in sample survey theory: an appraisal. *Canadian Journal of Statistics*. 1997; 25:1–21.
2. Cassel CM, Särndal CE, Wretman JH. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*. 1976; 63:615–620.
3. Särndal, CE.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling*. Springer-Verlag; New York: 1992.
4. Sugden RA, Smith TMF. Design-based properties of linear calibrated estimators of a finite population total. *International Statistical Review*. 2007; 75:218–223.
5. Royall, RM. The prediction approach to sampling theory. In: Krishnaiah, PR.; Rao, CR., editors. *Handbook of Statistics Volume 6: Sampling*. North-Holland/Elsevier; Amsterdam; New York: 1988. p. 399–413.
6. Hartley HO, Rao JNK. A new estimation theory of sample surveys. *Biometrika*. 1968; 55:547–557.
7. Robinson J. Conditioning ratio estimates under simple random sampling (C/R: V84 P352). *Journal of the American Statistical Association*. 1987; 82:826–831.
8. Holt D, Smith TMF. Post stratification. *Journal of the Royal Statistical Society, Series A (General)*. 1979; 142:33–46.
9. Stanek EJ III, Singer JM, Lencina VB. A unified approach to estimation and prediction under simple random sampling. *Journal of Statistical Planning and Inference*. 2004; 121:325–338.
10. Stanek EJ III, Singer JM. Predicting random effects from finite population clustered samples with response error. *Journal of the American Statistical Association*. 2004; 99:1119–1130.

11. Cochran, WG. Survey Sampling. 3. John Wiley and Sons; New York: 1977.
12. Kempthorne, O. New Developments in Survey Sampling. John Wiley and Sons Inter-Science; 1969. Some remarks on statistical inference in finite sampling; p. 358-389.
13. Rao JNK, Bellhouse DR. Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*. 1978; 2:125-141.
14. Rao TJ. Some aspects of random permutation models in finite population sampling theory. *Metrika*. 1984; 31:25-32.
15. Rao, CR. Foundations of Statistical Inference. Some aspects of statistical inference in problems of sampling from finite populations. Godambe, VP.; Sprott, DA., editors. Holt, Rinehart and Winston; Toronto: 1971. p. 177-202.
16. Royall RM. The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*. 1976; 71:657-664.
17. Li, W. PhD diss. University of Massachusetts; 2003. Use of random permutation model in rate estimation and standardization.
18. Pratt CA, Lemon SC, Fernandez ID, Goetzel R, Beresford SA, French SA, Stevens VJ, Vogt TM, Webber LS. Design characteristics of worksite environmental interventions for obesity prevention. *Obesity (Silver Spring)*. 2007; 15:2171-80. [PubMed: 17890484]
19. Montanari GE. Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*. 1987; 55:191-202.
20. Valliant, R.; Dorfman, AH.; Royall, RM. Finite Population Sampling and Inference, a Prediction Approach. John Wiley & Sons; New York: 2000.
21. Royall, RM. The prediction approach to finite population sampling theory: application to the hospital discharge survey. National Center for Health Statistics, Office of Statistical Methods; Rockville, MD: 1973. p. 1-31.

Appendix I

Under simple random sampling, the unbiasedness of the predictor for $\mathbf{C}'_{II}\mathbf{Z}_{II}$ requires $E(\mathbf{w}'\mathbf{Z}_I)=E(\mathbf{C}'_{II}\mathbf{Z}_{II})$, which implies $\mathbf{w}'\mathbf{G}_I - \mathbf{C}'_{II}\mathbf{G}_{II}=0$. The variance of the estimator for τ is thus $\text{var}(\hat{T}_y)=\mathbf{w}'\mathbf{V}_I\mathbf{w} - 2\mathbf{w}'\mathbf{V}_{I,II}\mathbf{C}_{II}+\mathbf{C}'_{II}\mathbf{V}_{II}\mathbf{C}_{II}$. Minimizing MSE of $\hat{T}_y$ is equivalent to minimizing $\text{var}(\hat{T}_y)$ subject to the unbiased constraint or alternatively, the following function,

$$\Phi(\mathbf{w})=\mathbf{w}'\mathbf{V}_I\mathbf{w} - 2\mathbf{w}'\mathbf{V}_{I,II}\mathbf{C}_{II}+2\{\mathbf{w}'\mathbf{G}_I - \mathbf{C}'_{II}\mathbf{G}_{II}\}\lambda,$$

where λ is a Lagrangian multiplier. Minimizing $\Phi(\mathbf{w})$ and applying Royall's prediction theorem [16], we obtain

$$\hat{\mathbf{w}}'=\mathbf{V}_I^{-1}\left\{\mathbf{V}_{I,II}+\mathbf{G}_I(\mathbf{G}'_I\mathbf{V}_I^{-1}\mathbf{G}_I)^{-1}(\mathbf{G}'_{II}-\mathbf{G}'_I\mathbf{V}_I^{-1}\mathbf{V}_{I,II})\right\}\mathbf{C}_{II},$$

which can be simplified to, $\hat{\mathbf{w}}'=N\bar{c}_{II}\left((1-f)\frac{\mathbf{1}'_n}{n}-\boldsymbol{\beta}'\otimes\frac{\mathbf{1}'_n}{n}\right)$.

Under stratified simple random sampling, the unbiasedness constraint for τ requires that

$\left(\mathbf{w}'\left(\bigoplus_{h=1}^H\left(\begin{pmatrix} 1 \\ \mathbf{0}_{Q\times 1} \end{pmatrix}\otimes\frac{\mathbf{1}_{n_h}}{N_h}\right)\right)-\mathbf{1}'_H\left(\bigoplus_{h=1}^H(1-f_h)\right)\right)=\mathbf{0}$. Derivation of the BLUE is similar to the above.

Table 1

Sampling Design of a Health Survey among Hospital Workers

Stratum (<i>h</i>)	Population Data			Sample		
	Number of workers	Mean age	Sampling fraction	Sample size	Mean age	
	N_h	μ_{hx}	f_h	n_h	$\bar{X}_h$	
White women (1)	1,799	43.4	0.043	77	44.2	
White men (2)	461	43.5	0.139	64	44.5	
Minority women (3)	208	38.8	0.288	60	39.8	
Minority men (4)	85	42.3	0.329	28	39.0	
Overall	2553	43.1	0.090	229	42.5	

Table 2
Crude and Adjusted Estimates of Fruit and Vegetable Consumption Index by Stratum and Overall

Stratum (<i>h</i>)	Crude estimate		Adjustment for age				Adjusted estimate	
	f_h	$\bar{Y}_{hI}$	$SE(\bar{Y}_{hI})$	$\hat{\beta}_h$	$X_{hII} - \mu_{ex1}$	$\hat{\beta}_{h1} (X_{hII} - \mu_{ex1})$	$\hat{T}_h$	$SE(\hat{T}_h)$
White women (1)	0.043	11.53	0.530	-0.030	0.76	-0.022	11.55	0.529
White men (2)	0.139	13.11	0.569	0.035	1.01	0.035	13.07	0.568
Minority women (3)	0.288	10.27	0.552	0.074	1.02	0.075	10.19	0.545
Minority men (4)	0.329	10.32	0.937	0.165	-3.33	-0.551	10.87	0.908
Overall	0.090	11.67	0.391	0.056	-0.61	-0.034	11.69	0.390