USE OF AUXILIARY INFORMATION
IN A DESIGN-BASED RANDOM PERMUTATION MODEL ¶

Wenjun Li §
Division of Preventative and Behavioral Medicine
University of Massachusetts Medical School, Worcester, MA 01655

Edward J. Stanek III
Department of Biostatistics and Epidemiology
University of Massachusetts, Amherst, MA 01003

ABSTRACT

This paper illustrates that prediction techniques, commonly used in model-based approach, can be applied to incorporating auxiliary information in design-based estimation. Design-based estimators of population parameters (such as mean or total) under simple random sampling without replacement using auxiliary information are developed by extending the design-based prediction method of Stanek, Singer and Lencina (2004) and Stanek and Singer (2004). The sampling is represented with indicator random variables, and the joint permutation of the response and auxiliary variables is represented using a set of simultaneous permutation equations. Known auxiliary information is incorporated through centering the auxiliary variables

¶ The content of this article is a part of the first author’s dissertation conducted at the Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts. This research was partially supported by a NIH grant (NIH-PHS-R01-HD36848). The authors wish to thank Drs. Julio Singer, John Buonaccorsi, and Carol Bigelow for their constructive comments.
§ Corresponding author, email address: Wenjun.Li@UMASSMED.EDU
on their respective known means. The derived estimators, defined as linear functions of the sample, are unbiased and have minimum mean squared error. They have functional forms similar to model-assisted and calibration estimators. Unlike these estimators, they are developed using optimality criteria identical to those used in a model-based prediction approach.

**Key words:** auxiliary variable, design-based inference, prediction, finite sampling random permutation model, simultaneous permutation

**Running title:** Use of auxiliary information
1. INTRODUCTION

Estimators can be made more precise by accounting for auxiliary information such as gender, age, income and chronic disease-bearing history that are partially or completely known in a population. Methods of improving estimation with auxiliary information have been discussed in both model-based (Bolfarine and Zacks 1992; Valliant et al. 2000) and design-based (Cassel et al. 1977; Cochran 1977; Särndal and Wright 1984; Deville and Särndal 1992; Särndal et al. 1992) approaches. Model based approaches include….. Design based approaches include .... However, the underlying conceptual frameworks differ between the two approaches.

The design-based approach is based on a probability model that arises from applying a sampling design to an identifiable finite population. Parameters are defined in the population, and optimal (unbiased) estimators minimize the expected mean squared error (MSE), with expectation taken over all possible samples (Horvitz and Thompson 1952; Cassel et al. 1977).

The model-based approach is based on a probability model that is postulated for a set of random variables in a superpopulation (Rao 1997; Brewer 1999; Valliant et al. 2000). An identifiable population is not defined; instead a realization of a set of superpopulation random variables stands for the ‘population’. Parameters are defined in the superpopulation (not the population), and linear combinations of random variables are the target of prediction, with bias, variance and mean squared error of the predictor defined in terms of the expectation over all possible realizations of a model for the superpopulation (Brewer 1995). The actual sample design plays no role in the inference.

The necessity of defining a superpopulation, and the lack of an identifiable population is a limitation of the model-based approach. However, this approach has been more readily adapted to complex analysis needs, extending many methods in traditional statistics. We develop a new method for accounting for auxiliary variables in a design based framework. The method is based on a random permutation probability model that represents a without replacement simple random sample (SRS) from a finite population. Parameters are clearly defined in terms of linear
functions of population units. To estimate a parameter, we representing it as a linear combination of sample and remainder random variables, we predict the linear combination of the remainder random variables borrowing methods from the model based approach. The results provide a solid theoretical bridge between the two frameworks.

Survey statisticians have made efforts to bridge the prediction- and design-based approaches (are there pure 'design based' and 'model based' approaches?), and to find estimators that have good properties and are interpretable within both approaches (Särndal and Wright 1984; Brewer et al. 1988; Deville and Särndal 1992; Särndal et al. 1992; Brewer 1995; Brewer 1999; 1999; Rao 1999). There are two subclasses of such methods for incorporating auxiliary information, namely, model-assisted (Cassel et al. 1976; Särndal et al. 1992) and calibration approaches (Deville and Särndal 1992). In a model-assisted approach, efficient estimators (are these model-based?) with good design-based properties are developed for a superpopulation model, with efficiency determined from a design based? probability sampling distribution (Deville and Särndal 1992). The generalized regression estimator (GREG) is an example of this type (Cassel et al. 1977; Särndal et al. 1992). In the calibration approach, a class of calibrated estimators is derived by applying a new set of weights that has minimum distance to the benchmark weights (such as the inverse of inclusion probabilities) while "calibrated" to known population quantities on some set of auxiliary variables. By forcing the new weights be close to initial weights, one wishes to preserve (or approximate) the unbiasedness and design-consistence properties of the Horvitz-Thompson estimators (Deville and Särndal 1992). The GREG and calibration approaches combine ideas from model based and design based approaches, but lack an integrated theoretical framework.

An integrated design based approach that takes advantage of prediction methods was developed by Stanek, Singer and Lencina (2004). Using a random permutation model representing SRS, they showed that the estimation of population parameters (such as the mean) can be accomplished by developing predictors of linear functions of non-sampled random variables.
variables using prediction techniques. Furthermore, Stanek and Singer (2004) extended this approach to prediction of random effects in balanced two-stage cluster sampling with response error. In this paper, we extend their method to incorporate auxiliary information. **Beginning with a vector of responses for each subject in a finite population, SRS of subjects is represented by a random permutation probability model that permutes subject response vectors.** Auxiliary information is incorporated through a simple linear transformation.

This paper is organized as follows. We **first** present definitions and notation, and introduce the random permutation model in the simplest scenario with one response and one auxiliary variable. We then define the population parameter of interest and derive the best linear unbiased estimator (BLUE). Subsequently, results are extended to scenarios with multiple auxiliary variables. We conclude by illustrating the method with an example, and include results from a small scale simulation evaluating empirical estimators. (Wenjun, I’d like to follow the notation of the JASA paper, not the JSPI paper).

2. DEFINITIONS AND NOTATION

Let the population consist of \( N \) subjects labeled \( s = 1, 2, \ldots, N \). We assume the labels are non-informative and serve only to identify the subjects. Associated with subject \( s \) is a non-stochastic potentially observable response vector, \( (y_s, x_s)' \), where \( y_s \) denotes the response of interest, and \( x_s = ((x_{is}) ) \) is a \( p \times 1 \) vector of auxiliary variables. We represent the population mean by the vector of means, \( (\mu_x, \mu_y)' \), where \( \mu_x = ((\mu_{ix}) ) \) is a \( p \times 1 \) vector. The population variance is defined by the \( (p+1) \times (p+1) \) matrix \( \frac{N-1}{N} \Sigma \), where

\[
\Sigma = \begin{pmatrix}
\begin{pmatrix}
\sigma_{y}^2 \\
\sigma_{yx}^2
\end{pmatrix} & \begin{pmatrix}
\sigma_{yx}^2 \\
\sigma_{yx}^2
\end{pmatrix}
\end{pmatrix},
\]

\( \sigma_{yx} = \begin{pmatrix}
\sigma_{y1} & \sigma_{y2} & \cdots & \sigma_{yp}
\end{pmatrix}' \) and \( \Sigma_s = \begin{pmatrix}
\sigma_{ys, s}\end{pmatrix}' \) is a \( p \times p \) matrix. Individual terms are...
defined in the usual manner, such that \( \mu_s = \frac{1}{N} \sum_{i=1}^{N} y_i \), \( \mu_x = \frac{1}{N} \sum_{i=1}^{N} x_{i} \), \( \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_y)^2 \), \( \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \) and \( \sigma_{yx} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_y)(x_i - \mu_x) \).

3. SAMPLING AND THE RANDOM PERMUTATION MODEL

We define random variables representing a simple random sample of size \( n \) from a population as the first \( n \) variates in a random permutation of subjects. We assume that all \( N \) subjects are randomly permuted with equal probability assigned to each possible permutation. The random variables representing a permutation define the random permutation model.

Following Stanek, Singer and Lencina (2004), we explicitly represent these random variables in terms of a set of indicator random variables \( U_i, i = 1, \ldots, N \), that have a value of one if subject \( s \) is in position \( i \) in a permutation, and zero otherwise. Using this notation, response for the subject in position \( i \) in a permutation is represented by the random variable \( Y_i = \sum_{s=1}^{N} U_i y_s \). Using vectors, \( Y = U' y \) where \( U = (U_1 \quad U_2 \quad \ldots \quad U_N)' \) represents a vector of random variables that selects a subject in position \( i \) and \( y = (y_i) \) represents the population response vector.

Defining \( U_{N \times N} = (U_1 \quad U_2 \quad \ldots \quad U_N)' \), the random variables representing a permutation of response are given by \( Y = U y \). Similar vectors can be defined to represent a permutation of auxiliary variables, \( X_k = U_k, k = 1, \ldots, p \). The vector and matrix notation is valuable to simplify representation of the problem and the development of estimators.

We use properties of the indicator random variables to evaluate the expected value and variance of the random variables. Taking expectation over all possible permutations,

\[ E(U_i) = \frac{1}{N} \text{ for all } i = 1, \ldots, N; s = 1, \ldots, N. \]

As a result, \( E(Y_i) = \mu_y \) for all \( i = 1, \ldots, N \). In a similar
manner, \( E(U_i U_{i'}) = \frac{1}{N(N-1)} \) when \( i \neq i' \) and \( s \neq s' \), \( E(U_i^2) = \frac{1}{N} \) when \( i = i' \) and \( s = s' \), and \( E(U_i U_{i'}) = 0 \) otherwise. Using these expressions, we can show that \( \text{var}(Y_i) = \frac{N-1}{N} \sigma_y^2 \) for all \( i = 1, \ldots, N \), and that \( \text{cov}(Y_i, Y_{i'}) = -\frac{1}{N} \sigma_y^2 \) when \( i \neq i' \).  

We combine these results to express the expected value and variance of a random permutation using vector notation. As a result, \( E(Y) = \mu_x I_N \) where \( I_N \) is an \( N \times 1 \) column vector; and \( \text{var}(Y) = \sigma_y^2 P_N \), where \( P_N = I_N - \frac{1}{N} J_N \), \( J_N \) is an \( N \times N \) identity matrix, and 

\[ J_N = I_N I_N \].  Similar expressions apply to auxiliary variates.

4. SIMULTANEOUS RANDOM PERMUTATION MODEL

We next represent random variables arising from the random permutation of response and auxiliary variates simultaneously. For simplicity, (in Section 4-7) suppose that the response variable is an indicator of a subject’s cigarette smoking status, and that there is a single auxiliary variable (i.e. \( p = 1 \)) corresponding to the subject’s age (which we represent by \( x_i \)). We assume that age is known for all subjects in the population.

We define the simultaneous random permutation model by concatenating the permutation vectors for response and the auxiliary variates. The resulting model is similar to the seemingly unrelated regression model (Zellner 1963) such that

\[
\begin{bmatrix}
Y \\
X
\end{bmatrix}
= \begin{bmatrix}
I_N & 0 \\
0 & I_N
\end{bmatrix}
\begin{bmatrix}
\mu_x \\
\mu_y
\end{bmatrix} + E
\]

\[ Z = G\mu + E \]  

where \( Z = (Y' \ X')' \), \( G = I_2 \otimes I_N \), and \( \mu = (\mu_x \ \mu_y)' \). In this model, \( E(Z) = G\mu \) and \( \text{var}(Z) = \Sigma \otimes P_N \).
The mean age, \( \mu_i \), is known since age is assumed to be known for all subjects. This implies that the \( N \) random variables arising from permuting the auxiliary variable are constrained by \( \frac{1}{N} \sum_{i=1}^{N} x_i = \mu_i \). We eliminate \( \mu_i \) from model (1) by subtracting \( \frac{1}{N} \sum_{i=1}^{N} x_i \) from both sides (which is equivalent to multiplying each term in the model by \( \frac{1}{N} I_N \) ).

Representing the elements of \( X' \) by \( x_i = x_i - \mu_i \), the transformed model is

\[
\begin{bmatrix}
Y \\
X'
\end{bmatrix} = \begin{bmatrix} 1_{N \times 1} \\
0_{N \times 1}
\end{bmatrix} \mu_x + E'
\]

(2)

where \( Z' = RZ \) and \( G' = RG \). Since \( P_N \) is idempotent, \( E(Z') = G' \mu_x \), \( \text{cov}(Z') = \Sigma \otimes P_N \).

[The design matrix structure is similar to the design matrices described for growth curve models by] Rao (1967). Put in discussion

5. SAMPLING AND PARTITIONING THE RANDOM MATRICES

We represent random variables for a SRS of size \( n \) by the first \( n \) elements of the random permutation vectors. The sampled \( (Z_i') \) and remaining \( (Z_{ii}') \) portions of \( Z' \) can be obtained by pre-multiplication by \( K = \left[ \begin{array}{cc}
I_2 \otimes (I_n \ 0_{n \times (N-n)}) \\
I_2 \otimes (0_{(N-n) \times n} \ I_{N-n})
\end{array} \right] \), i.e., \( KZ' = (Z_i' \ Z_{ii}') \). Since \( K \) is non-stochastic, it follows that \( E(Z_i') = \left( \mu_i, 1_n \right)' \), \( E(Z_{ii}') = \left( \mu_{N-n}, 0 \right)' \), and

\[
\text{cov} \begin{bmatrix} Z_i' \\
Z_{ii}'
\end{bmatrix} = \begin{bmatrix} V_j & V_{i,ii} \\
V_{i,ii} & V_{ii}
\end{bmatrix}
\]

where \( V_j = \Sigma \otimes P_{n,N} \), \( V_{i,ii} = \Sigma \otimes P_{(N-n),N} \) and
partitioned model that reflects SRS sampling can be represented as

\[
\begin{pmatrix}
\mathbf{Z}'_i \\
\mathbf{Z}''_i
\end{pmatrix} = \begin{pmatrix}
\mathbf{G}'_i \\
\mathbf{G}''_i
\end{pmatrix} \mu' + \mathbf{E}',
\]

(1)

where \( \mathbf{G}'_i = \left( \mathbf{1}'_a, 0_{a \times (N-a)} \right)' \) and \( \mathbf{G}''_i = \left( \mathbf{1}'_{N-n}, 0_{(N-n) \times a} \right)' \).

6. PARAMETER OF INTEREST

This is problematic. The coefficients \( c_i \) depend on position, so I can’t think of a rational for any coefficient other than \( c_i = k \) for all \( i = 1, \ldots, N \). If \( c_i \) can be some other values, then \( \sum_{i=1}^{N} c_i Y_i \) may be a random variable. If it is a random variable, then we would “predict” it, not ‘estimate’ it. I think we should focus on estimating the population mean (not the total), in which case we don’t have to introduce new notation. This would also imply that \( c_i = k \). What do you think? We assume that the parameter of interest is a linear function of the permuted random variables, namely, \( T_j = \sum_{i=1}^{N} c_i Y_i = \sum_{i=1}^{a} c_i Y_i + \sum_{i=a+1}^{N} c_i Y_i \), or in matrix format, \( (I \text{ don't like the notation } T, \text{ since this represents a random variable}) \)

\[
T_j = \mathbf{C}' \mathbf{Z}' = \mathbf{C}'_1 \mathbf{Z}'_1 + \mathbf{C}'_2 \mathbf{Z}'_2,
\]

(2)

where \( \mathbf{C} = \left( \mathbf{c}' \ 0_{N \times a} \right)' \), \( \mathbf{C}_1 = \left( \mathbf{c}'_1 \ 0_{a \times 1} \right)' \) and \( \mathbf{C}_2 = \left( \mathbf{c}'_2 \ 0_{(N-a) \times 1} \right)' \), \( \mathbf{c} = (c_1 \ c_2 \ \cdots \ c_N)' \), \( \mathbf{c}_1 = (c_1 \ c_2 \ \cdots \ c_a)' \) and \( \mathbf{c}_2 = (c_{a+1} \ c_{a+2} \ \cdots \ c_N)' \) denoting vectors of known constants. If \( c_i = 1 \), \( i = 1, 2, \ldots, N \), then \( T_j \) is the population total of the response variate. After sampling, only the second term in the right hand side of (2) will be unknown; thus, estimating \( T_j \), is equivalent to
find a predictor of $C'_H Z'_H = \sum_{i=1}^{N} c_i Y_i$. We develop the best linear unbiased predictor (BLUP) of $C'_H Z'_H$, and refer to the estimator of $\mu_j$ as the BLUE of $\mu_j$.

7. BLUE OF A LINEAR FUNCTION

We require the estimator of $\sum_{i=1}^{N} c_i Y_i$ to be a linear function of the sample,

$$w'Z'_i = \sum_{i=1}^{n} w_y Y_i + \sum_{i=1}^{n} w_x X'_i,$$

where $w = (w_y \quad w_x)'$ is a $2n \times 1$ vector of coefficients,

$$w_y = (w_{y1} \quad w_{y2} \cdots \quad w_{yn})'$$
and $w_x = (w_{x1} \quad w_{x2} \cdots \quad w_{xn})'$, to be unbiased, i.e.,

$$E\left(w'Z'_i\right) = E\left(C'_H Z'_H\right),$$

and have minimum MSE. The estimator of $T_j$ is given by

$$P = (C'_j + w')Z'_i = \sum_{i=1}^{n} w'_y Y_i + \sum_{i=1}^{n} w'_x X'_i.$$  (3)

The unbiased constraint requires that $w'G'_j - C'_H G'_{H,i} = 0$. The variance of $P$ is

$$\text{var}(P) = w'V_j w - 2w'V'_{j,i} C'_{Hi} + C'_{Hi} V_{Hi}.\quad \text{With such setup, the prediction theorem of Royall (1976) can be applied.}$$

The minimum variance unbiased estimator is obtained by minimizing the function

$$\Phi(w) = w'V_j w - 2w'V'_{j,i} C'_{Hi} + 2\left(w'G'_j - C'_H G'_{H,i}\right) \lambda,$$  (4)

where $\lambda$ is a Lagrangian multiplier, resulting in

$$\dot{w} = V_j^{-1} \left[V_{j,i} + G'_j \left(G'_{j,i} V_{j,i}^{-1} G'_j\right)^{-1} \left(G'_{H,i} - G'_{j,i} V_{j,i}^{-1} V_{j,i} G'_{H,i}\right)\right] C'_{Hi}.$$  (5)

After simplification, (5) reduces to

$$\hat{w}_{ji} = \left(\frac{1}{N-n} \sum_{i=1}^{n} c_i\right) \left(\frac{1}{f} \right) \left(1-f \right) \beta_i, \ i = 1, 2, \ldots, n.$$  (6)

where $f = n/N$, $\beta = \sigma_{y{i}}/\sigma_{y}^2$. Consequently, the estimator and its variance are

$\dot{w}_{ji}$ and $\text{var}(\hat{w}_{ji})$. 

\[\text{auxiliary_simplified_2005-ejs.doc} \quad 06/29/05\]
\[
P = \sum_{i=1}^{n} c_i Y_i + \sum_{i=1}^{N} \left\{ c_i \left( \bar{Y}_i - \frac{N}{N-n} \beta (\bar{X}_i - \mu_x) \right) \right\}
\]

and

\[
\text{var}(\hat{P}) = \left( \sum_{i=1}^{n} c_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} c_i \right)^2 \right) \sigma_y^2 + n(1-f) \left( \frac{1}{n} \sum_{i=1}^{n} c_i - \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right)^2 \sigma_y^2
\]
\[+2N(1-f) \left( \frac{1}{n} \sum_{i=1}^{n} c_i - \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right) \left( \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right) (1-\rho^2) \sigma_y^2
\]
\[+N^2 \left( \frac{1-f}{n} \right) \left( \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right)^2 (1-\rho^2) \sigma_y^2,
\]  

where \( \rho = \sigma_{xy}/\sigma_y \) is the correlation coefficient of \( Y \) on \( X \).

The estimator of a population total \( N \mu_y \) may be obtained as an application of (7) and (8) by setting \( c_i = 1, \ i = 1,2,\ldots,N \). It follows that \( \hat{Y}_i = \frac{1}{f} \left( 1-f \right) \beta, \ i = 1,2,\ldots,n \), and that,

\[
\hat{T}_y = n \bar{Y}_f + (N-n) \left[ \bar{Y}_f - \frac{N}{N-n} \beta (\bar{X}_f - \mu_x) \right],
\]

\[
\text{var}(\hat{T}_y) = (1-\rho^2) (N^2 \left( \frac{1-f}{n} \right) \sigma_y^2),
\]

where \( \bar{Y}_f \) is the sample mean of the response variable and \( \bar{X}_f \) is the sample mean of the auxiliary variable. This expression is similar to the expressions commonly seen in linear regression estimators, but includes a finite population correction factor. \( \text{(This needs re-writing if this is the only parameter considered.)} \)

The estimators derived in the previous section depend on the covariance \( \sigma_{xy} \), and on the variance \( \sigma_x^2 \). While \( \sigma_x^2 \) may be known, \( \sigma_{xy} \) is often unknown in practice. A simple commonly

\[
\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_f)^2, \quad \hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_f)(X_i - \bar{X}_f), \quad \text{and} \quad \hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_f)^2.
\]

auxiliary_simplified_2005-ejs.doc
where $\bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij}$ and $\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}$. An approximate normal-based interval estimator for $\hat{P}$ is given by $\hat{P} \pm z_{1-\alpha} \sqrt{\text{var} (\hat{P})}$, where $\hat{P}$ is defined by replacing variance components in (8) or (10) by their estimates.

[Are you going to discuss what happens when the auxiliary means are not known?]

8. EXTENSION TO CASES WITH MULTIPLE COVARIATES

(With the previous changes, this section can be shortened. I haven’t done so.) Extension of the above results to scenarios with multiple auxiliary variables is straightforward. Suppose $p$ auxiliary variables are used in estimation. The matrix of random variables representing a joint permutation of a response variable and the $p$ covariates is given by

$$U(y, x_1, \ldots, x_p) = (Y, X_1, \ldots, X_p),$$

where $X_i$ is the random vector corresponding to the $k$th covariate, $k = 1, 2, \ldots, p$. A simultaneous random permutation model can be defined similar to

$$E(Z) = G\mu,$$

and that $\text{cov}(Z) = \Sigma \otimes P_N$, where $\Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \Sigma_X \end{pmatrix}$, $\sigma_{xy}$ is a column vector of covariance between $Y$ and $X = (X_1, X_2, \ldots, X_p)$, and $\Sigma_X$ is the variance-covariance matrix of $X$. When auxiliary means $\mu_k$, $k = 1, \ldots, p$, are known, similar transformation can be applied to $Z$, such that $Z^* = \begin{pmatrix} I_N & 0 \\ 0 & I_p \otimes P_N \end{pmatrix} Z$. A reparameterized model incorporating known auxiliary information has a form identical to Model , but here $G^* = \begin{pmatrix} I_N & 0'_{N \times (p-1)} \end{pmatrix}$. With these, linear unbiased minimum variance estimation can be derived following the same steps in Section 7.
Detailed derivation is given in the Appendix. The unique solution for the vector of coefficients $\mathbf{w}$ is

$$
\hat{\mathbf{w}} = \left( \frac{1}{N-n} \sum_{i=1}^{n} c_i \right) \left[ \frac{1}{f} \left( \frac{1-f}{\mathbf{1}_n} \right) \otimes \mathbf{1}_n \right],
$$

where $\mathbf{\beta} = \Sigma^{-1}_x \mathbf{\sigma}_{xy} = \left( \beta_1, \beta_2, \ldots, \beta_p \right)'$ with $\beta_k = \sigma_{yk} / \sigma_y^2$, $k = 1, 2, \ldots, p$. I don't understand this. The diagonal element of $\Sigma$ is $\sigma_i^2$. The diagonal element of $\Sigma^{-1}_x$ is not $\sigma_i^2$ unless $\Sigma$ is diagonal. ?? $\mathbf{\beta} = \Sigma^{-1}_x \mathbf{\sigma}_{xy} = \left( \beta_1, \beta_2, \ldots, \beta_p \right)'$ Accordingly, the estimator and its variance are

$$
\hat{P} = \sum_{i=1}^{n} c_i \bar{Y}_i + \sum_{i=1}^{n} \left\{ c_i \left( \bar{Y}_i - \frac{N}{N-n} \sum_{i=1}^{n} \beta_k \left( \bar{X}_{ik} - \mu_k \right) \right) \right\}
$$

and

$$
\text{var}(\hat{P}) = \left( \sum_{i=1}^{n} c_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} c_i \right)^2 \right) \sigma_y^2 + n(1-f) \left( \frac{1}{n} \sum_{i=1}^{n} c_i - \frac{1}{N-n} \sum_{i=1}^{n} c_i \right)^2 \sigma_y^2
$$

$$
+ 2N(1-f) \left( \frac{1}{n} \sum_{i=1}^{n} c_i - \frac{1}{N-n} \sum_{i=1}^{n} c_i \right) \left( \frac{1}{N-n} \sum_{i=1}^{n} c_i \right) \left( 1 - \rho^2_{yk} \right) \sigma_y^2
$$

$$
+ N^2 \frac{(1-f)}{n} \left( \frac{1}{N-n} \sum_{i=1}^{n} c_i \right)^2 \left( 1 - \rho^2_{yk} \right) \sigma_y^2,
$$

where $\bar{Y}_i$ is the sample mean of the response variable and $\bar{X}_{ik}$ is the sample mean of the $k$-th auxiliary variable, and $\rho^2_{yk} = \sigma_{yk}^2 \Sigma_{xx}^{-1} \mathbf{\sigma}_{xy} / \sigma_y^2$ is the multiple correlation coefficient of $Y$ on $X'$. Notice that the only difference between (12) and (8) is the multiple correlation coefficient $\rho^2_{yk}$ instead of $\rho^2$. A straightforward application of (11) and (12) is to obtain estimator of population total $T_y$ by letting $c_i = 1$, $i = 1, 2, \ldots, N$, that is,
\( \hat{T}_y = n \bar{y} + (N-n) \left[ \bar{Y}_y - \frac{N}{N-n} \sum_{i=1}^{n} \beta_i \left( \bar{X}_i - \mu_{x_i} \right) \right], \) \hfill (13)

\[ \text{var} \left( \hat{T}_y \right) = (1 - \rho_{x\varepsilon}^2) \left( N^2 \left( \frac{1-f}{n} \right) \sigma_y^2 \right). \] \hfill (14)

This expression is similar to the expressions commonly seen in multiple linear regression models (Graybill 1976), but includes a finite population correction factor. Results (13) and (14) are also the same as difference estimators with optimal coefficients (Montanari 1987).

[Somewhere there should be the relationship with the GREG and calibration approaches.]

9. APPLICATION EXAMPLES

In this section, we apply the BLUE to an example related to the estimation of the total number of smokers in a finite population of size \( N \). The smoking status \( y \) is coded as 1 for smoker and 0 otherwise. The age of the subjects is denoted by \( x \), and could be continuous or dichotomous (or categorical), e.g., 0 for younger (<65) or 1 for older (\( \geq 65 \)). When age is continuous, \( \mu_x \) is the population mean age; when age is dichotomous, \( \mu_x \) is its proportion of persons aged \( \geq 65 \). In either case, we assume \( \mu_x \) is known. The estimator \( \hat{P} \) and its variance constitute simple examples of (9) and (10). Furthermore, \( \beta = \sigma_{xy} / \sigma_x^2 \) is the population regression coefficient of smoking status on age or age categories, \( \bar{X}_y \) is the mean age (or proportion aged \( \geq 65 \)) in the sample, and \( \rho^2 = \sigma_{xY}^2 / \sigma_x^2 \sigma_y^2 \) is the squared population correlation coefficient between smoking status and age.

When the auxiliary variable is dichotomous, i.e., the proportion of subjects \( \geq 65 \) is known, the total numbers of persons aged <65 and \( \geq 65 \) in the population (\( N_a \) and \( N_b \), respectively) are also known, and \( \sigma_Y^2 = N_a N_b / N (N-1) \). When \( \sigma_Y^2 \) and \( \sigma_{xy} \) are unknown, we replace \( \sigma_Y^2 \) and
\( \sigma_{y} \), with their sample estimates, \( \hat{\sigma}^2_y = \frac{m(n-m)}{n(n-1)} \) and \( \hat{\sigma}^2_{xy} = \frac{n_{a}n_{b}}{n(n-1)} \left( \frac{m_a - m_b}{n_a - n_b} \right) \), where \( n_a \) and 
\( n_b \) are numbers of the younger and older in the sample, \( m = m_a + m_b \), \( m_a \) and \( m_b \) are the numbers of the younger and the older smokers in the sample, respectively. As a result, \( \hat{\beta} = \frac{N(N-1)}{n(n-1)} \left( \frac{n_{a}n_{b}}{N_aN_b} \right) \left( \frac{m_a - m_b}{n_a - n_b} \right) \). Consequently,

\[
\hat{\mu}_y = \frac{m}{n} + \frac{N-1}{n-1} \left( \frac{n_{a}n_{b}}{N_aN_b} \right) \left( \frac{m_a - m_b}{n_a - n_b} \right)
\]

and its variance is

\[
V(\hat{\mu}_y) = \frac{1}{n} \left[ m \left( \frac{m_a^2 - m_b^2}{n_a n_b} \right) \right].
\]

If we replace \( \sigma^2_y, \sigma_{xy} \) and \( \sigma^2_{xy} \) with their sample estimates \( \hat{\sigma}^2_y, \hat{\sigma}_{xy} \) and \( \hat{\sigma}^2_{xy} = \frac{n_{a}n_{b}}{n(n-1)} \), then \( \hat{\beta} = \frac{m_a}{n_a} - \frac{m_b}{n_b} \), \( \hat{\mu}_y = \frac{N}{n} \left( \frac{m_a}{n_a} \right) + \frac{N}{n} \left( \frac{m_b}{n_b} \right) \), and its estimated variance is

\[
V(\hat{\mu}_y) = \frac{1}{n} \left[ m \left( \frac{m_a^2 + m_b^2}{n_a n_b} \right) \right].
\]

This estimator is a “post-stratified” estimator of prevalence rate, which has been considered in calibration (Deville and Särndal 1992) and model-assisted approaches (Särndal et al. 1992).

Extension to scenarios where age is classified in multiple categories is straightforward.

When population covariances of the auxiliary variables are used, the estimator is

\[
\hat{\mu}_y = \frac{m}{n} - \frac{N-1}{n(n-1)} \sum_{k=1}^{K} \left( \frac{n_{k}}{N_k} \left( \frac{m_{k} - m_{k}}{n_{k}} \right) \right),
\]

and its variance is

\[
\text{var}(\hat{\mu}_y) = \left( \frac{1}{n} \right) \left( \frac{m(n-m)}{n(n-1)} - \frac{N-1}{n^2(n-1)^2} \sum_{k=1}^{K} \frac{1}{N_k} (mm_k - mn_k)^2 \right).
\]

When sample variance of the
auxiliary variables is used, the BLUE of $\mu_y$ is

$$\hat{\mu}_y = \sum_{k=1}^{N} \frac{N_k}{N} m_k,$$

and its variance is

$$\text{var}(\hat{\mu}_y) = \frac{1 - f}{n(n-1)} \left( m - \sum_{k=1}^{N} m_k \right).$$

(This section needs some work, rewriting with new simulation results)

Impact of choices of variance components on performance of the estimators

To understand the impact of estimating variance components on the performance of the estimators, we conducted a Monte Carlo simulation. A series of hypothetical populations of sizes 50, 100, 200, 400, 800 and 1,600 were generated; each with an auxiliary variable representing 50% men. In all populations, the prevalence of smokers is 60% for men and 20% for women. We then used (11) and (13) to estimate the population prevalence. The simulation study was carried out as follows. From each population of size $N = 50, 100, 200, 400, 800, 1600$, a series of 5,000 samples of size $n$, where $n = 25, 50, 100, 200, 400, 800$ and $n < N$, were drawn via SRS. For each sample, we estimated the proportion of smokers, and its variance using the following methods: 1) Simple Expansion Estimator without auxiliary information (SEE), 2) Design-based Estimator using Sample (DES) estimate of population variance of gender $\left( \hat{\sigma}_x^2 \right)$ and estimated covariance $\left( \hat{\sigma}_{xy} \right)$, 3) Design-based Estimator using known Population variance (DEP) of gender $\left( \sigma_x^2 \right)$ and sample estimates of $\left( \hat{\sigma}_{xy} \right)$, and 4) Design-based estimator using known Population variance and covariance (POP). The 95% confidence intervals (CIs) were computed using the variance estimators based on normal approximations. We then compared the overall performance of these estimation methods in terms of mean squared error (MSE), coverage rates.
and average length of the confidence intervals. Simulation results are summarized in Tables 1 and 2.

These simulations confirm that all these estimators are asymptotically unbiased, or at least that the biases are negligible; especially when the sample size is not small. There was a 12%-16% reduction in the MSE of the estimators relative to SEE for small and large sample sizes (Table 1). There is evidence that using known population auxiliary variances (DEP) is superior to using its sample estimates (DES) in terms of MSE reduction, especially when sample sizes are relatively small (n<100). The coverage rate of the confidence intervals constructed using the variance estimators of the four methods are all very close to their nominal level (95%), but the coverage rates of DEP and DES confidence intervals are lower than SEE and POP. The width of the DEP and DES confidence intervals tend to be the narrower than those of SEE and POP estimators. When sample sizes are moderate or large, there are little differences in coverage rates between the methods, or in the difference between DEP and POP methods.

10. DISCUSSION

We proposed an alternative method of developing estimators in the presence of auxiliary information using a random permutation model under SRS. This framework does not require parametric distributions. The random variables are attributable only to random sampling (permutation). We represented the joint permutation of response and auxiliary variables with a setup similar to seemingly unrelated regression. Best linear unbiased estimators are derived under design-based framework mimicking estimation techniques that are commonly applied in prediction-based approaches (Royall 1976; Valliant et al. 2000).

With this proposed method, we obtain explicit expressions for the joint permutation of response and multiple auxiliary variables, and formulae for the population mean vector and their variance-covariance matrix. These results are similar to those commonly seen in literature regarding seemingly unrelated regression (Zellner 1962; Zellner 1963; Revankar 1974; Revankar...
1976; Srivastava and Giles 1987; Gao and Huang 2000; Liu 2000), but our results incorporate
the finiteness of the population and do not rely on parametric or superpopulation model
assumptions.

The estimators presented in this article are identical to the estimators obtained under a
design-based approach (Cochran 1977) assuming a linear regression model or the difference
estimator (See Section 6.3 in (Särndal et al. 1992)) and GREG estimators derived under model-
assisted approach (Särndal et al. 1992). In fact, an appeal of the approach is providing a formal
foundation for these estimators. In addition, the estimators and their variances are the same as
for asymptotically design-unbiased estimators and their asymptotic design-variance derived from
a model-calibration approach based on a superpopulation framework (Wu and Sitter 2001).
However, our results depend neither on the required assumptions by superpopulation models
and model-assisted approaches nor on arbitrary distance measures between “adjusted” and
“initial” weights that are critical for the conventional calibration approach. In addition, the
estimators proposed here and their variances are exact and not approximated.

These estimators require the knowledge of both \(\sigma_{yX}\) and \(\Sigma_X\), and their variances are
always smaller than the variances of simple expansion estimators. The gain in the precision of
prediction is proportional to the squared (multiple) correlation coefficient of the response variable
and auxiliary variable(s). When both \(\sigma_{yX}\) and \(\Sigma_X\) are known, the derived estimator is the same
as difference estimator with optimal coefficients (Montanari 1987). When both \(\sigma_{yX}\) and \(\Sigma_X\) are
unknown, we apply these results by substituting their sample estimates (method of moments),
and the resulting estimators will be the same as usual calibration estimators.

The approach yields unbiased linear estimators whose weights are not sample-
dependent (when variance components are known). In contrast, usual calibration technique of
incorporating auxiliary information leads to weights that are conditioned on the observed sample,
and the corresponding estimators are not necessarily unbiased. The design-based random
permutation model estimators are optimized in terms of its mean squared error or its variance, a
direct and natural measure of precision; while calibration estimators are optimized in terms of the
distance between adjusted and naïve weights, with no assurance about minimum MSE or
variance. Another advantage of the estimators is that they are always unique when the
population variance (or variance-covariance matrix) of the auxiliary variable(s) is known and
nonsingular.

Finally, we acknowledge the limitations of this research. We considered only linear
unbiased minimum variance estimators under SRS and use of auxiliary variables at the
estimation stage. It remains to understand the theory and applications of this design-based
approach in other estimation/prediction problems, such as small-domain estimation, unit non-
response adjustment, and estimation of non-linear functions. It is interesting to examine whether
relaxing the unbiasedness constraints or using other optimal criteria such as log-determinants
will result in improvement or loss of efficiency in estimation. Another challenge is development of
estimation/prediction methods for more complex sampling design, such as stratified or cluster
random sampling and longitudinal survey sampling. We are currently making efforts toward these
directions, building on the results of these methods and Stanek and Singer (2004).
Reference


Table 1 Mean squared error (x 10,000) of four estimators

<table>
<thead>
<tr>
<th>N</th>
<th>SEE</th>
<th>DES</th>
<th>DEP</th>
<th>POP</th>
<th>SEE</th>
<th>DES</th>
<th>DEP</th>
<th>POP</th>
<th>SEE</th>
<th>DES</th>
<th>DEP</th>
<th>POP</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>48.7</td>
<td>43.1</td>
<td>43.0</td>
<td>40.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>73.9</td>
<td>65.6</td>
<td>65.3</td>
<td>62.4</td>
<td>24.3</td>
<td>20.9</td>
<td>20.9</td>
<td>20.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>85.5</td>
<td>75.1</td>
<td>74.8</td>
<td>71.7</td>
<td>36.0</td>
<td>31.0</td>
<td>31.0</td>
<td>30.2</td>
<td>12.2</td>
<td>10.3</td>
<td>10.3</td>
<td>10.1</td>
</tr>
<tr>
<td>400</td>
<td>89.9</td>
<td>79.2</td>
<td>78.8</td>
<td>75.6</td>
<td>42.0</td>
<td>36.1</td>
<td>36.0</td>
<td>35.3</td>
<td>17.9</td>
<td>15.1</td>
<td>15.0</td>
<td>14.9</td>
</tr>
<tr>
<td>800</td>
<td>92.9</td>
<td>81.5</td>
<td>81.1</td>
<td>78.1</td>
<td>45.9</td>
<td>39.2</td>
<td>39.1</td>
<td>38.4</td>
<td>21.1</td>
<td>17.7</td>
<td>17.7</td>
<td>17.5</td>
</tr>
<tr>
<td>1600</td>
<td>94.8</td>
<td>83.1</td>
<td>82.6</td>
<td>79.2</td>
<td>46.5</td>
<td>39.9</td>
<td>39.9</td>
<td>39.0</td>
<td>22.5</td>
<td>19.0</td>
<td>19.0</td>
<td>18.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>SEE</th>
<th>DES</th>
<th>DEP</th>
<th>POP</th>
<th>SEE</th>
<th>DES</th>
<th>DEP</th>
<th>POP</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>6.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>9.0</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>3.0</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>1600</td>
<td>10.4</td>
<td>8.8</td>
<td>8.8</td>
<td>8.7</td>
<td>4.5</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
</tr>
</tbody>
</table>

SEE  Simple Expansion Estimator  
DES  Design-based Estimator using the Sample Estimate of the Variance  
DEP  Design-based Estimator using the partially known Population Variance  
POP  Design-based Estimator using known Population Variance
Table 2 Coverage rates and width of 95% confidence intervals

<table>
<thead>
<tr>
<th>N</th>
<th>n=25</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEE</td>
<td>DES</td>
<td>DEP</td>
</tr>
<tr>
<td>50</td>
<td>95.8</td>
<td>93.0</td>
<td>93.1</td>
</tr>
<tr>
<td></td>
<td>93.1</td>
<td>92.1</td>
<td>92.1</td>
</tr>
<tr>
<td>100</td>
<td>91.2</td>
<td>92.6</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td>35.9</td>
<td>32.1</td>
<td>32.1</td>
</tr>
<tr>
<td>200</td>
<td>94.5</td>
<td>92.6</td>
<td>92.7</td>
</tr>
<tr>
<td></td>
<td>37.2</td>
<td>33.2</td>
<td>33.2</td>
</tr>
<tr>
<td>400</td>
<td>94.2</td>
<td>92.6</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td>37.8</td>
<td>33.6</td>
<td>33.6</td>
</tr>
<tr>
<td>800</td>
<td>93.9</td>
<td>92.5</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td>38.1</td>
<td>33.9</td>
<td>33.8</td>
</tr>
<tr>
<td>1600</td>
<td>94.7</td>
<td>94.7</td>
<td>94.7</td>
</tr>
<tr>
<td></td>
<td>94.4</td>
<td>94.7</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>94.8</td>
<td>94.7</td>
<td>94.7</td>
</tr>
<tr>
<td></td>
<td>94.7</td>
<td>94.6</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>94.1</td>
<td>94.7</td>
<td>94.7</td>
</tr>
</tbody>
</table>

SEE  Simple Expansion Estimator
DES  Design-based Estimator using the Sample Estimate of the Variance
DEP  Design-based Estimator using the partially known Population Variance
POP  Design-based Estimator using known Population Variance
Appendix 1  Proof of result (13) and (14)

When there are $p$ auxiliary variables, a simultaneous random permutation model can be defined similar to with $Z = \text{vec}(Y \ X_1 \ \cdots \ X_p)$ and $G = (I_{p+1} \otimes I_x)$. When auxiliary means

$\mu_k, k = 1, \ldots, p$, are known, let $X^* = P_nX$. The transformation can be represented as $Z^* = RZ$,

where $R = \begin{pmatrix} I_x & 0 \\ 0 & I_p \otimes P_N \end{pmatrix}$, and $RG = \begin{pmatrix} G^* & 0 \end{pmatrix}N(p+1)$. where $G^* = \begin{pmatrix} 1' x & 0_{N p+1} \end{pmatrix}'$, and $E^* = RE$.

The reparameterized model that incorporates the constraints thus has a form identical to Model .

To represent the SRS sampling, let $K = \begin{pmatrix} I_{p+1} \otimes \left( \begin{matrix} I_n & 0_{n(N-n)} \end{matrix} \right) \\ I_{p+1} \otimes \left( \begin{matrix} 0_{(N-n)n} & I_{N-n} \end{matrix} \right) \end{pmatrix}$, and thus the portioned simultaneous permutation model takes the same form of (1),

\[
\begin{pmatrix} Z'_r \\ Z_{II}' \end{pmatrix} = \begin{pmatrix} G'_r \\ G_{II}' \end{pmatrix} \mu_y + E',
\]

where $G'_r = \begin{pmatrix} 1' x & 0_{n p} \end{pmatrix}'$ and $G_{II}' = \begin{pmatrix} 1' n_{N-n} & 0_{1(N-n)p} \end{pmatrix}'$.

A linear function of the random variables can be defined similar to (2), with $C = \begin{pmatrix} c' & 0_{n p} \end{pmatrix}'$, $C_i = \begin{pmatrix} c_i' & 0_{n p i} \end{pmatrix}'$ and $C_{II} = \begin{pmatrix} c_{II}' & 0_{1(N-n)p} \end{pmatrix}'$. The estimator of $C'_r Z'_r$ can then be defined as a linear function of the sample, $w' Z'_r$, where $w = \begin{pmatrix} w_y' & w_r' \end{pmatrix}'$,

$w = \begin{pmatrix} w_1' & w_2' & \cdots & w_p' \end{pmatrix}'$, is a $(p+1) n x 1$ vector of coefficients. With these, a linear unbiased minimum variance estimation can be derived by minimizing the function

\[\Phi(w) = w' V_i w - 2w' V_i C_{II} + 2 \left( w' G'_r - C'_r G_{II}' \right) \lambda.\]
Specifically, differentiating the above function with respect to \( w \) and \( \lambda \), and setting the derivatives to zeros results in the following estimating equations

\[
\begin{pmatrix}
V_i & G_i^* \\
G_i'' & 0
\end{pmatrix}
\begin{pmatrix}
\dot{w} \\
\dot{\lambda}
\end{pmatrix} = \begin{pmatrix}
V_{i,t} C_H \\
G_{ii,t} C_H
\end{pmatrix},
\]

where \( G_i^* = \left( I_n' \ 0_{1 \times p} \right)' \) and \( G_H = \left( I'_{N-n} \ 0_{(N-n) \times p} \right)' \). The unique solution is

\[
\dot{w} = V_i^{-1}\left\{ V_{i,t} + G_i^* \left( G_i'' V_i^{-1} G_i^* \right) \left( G_{ii,t} - G_i'' V_i^{-1} V_{i,t} \right) \right\} C_H.
\]

(A.1)

Since \( V_i^{-1} = \Sigma^{-1} \otimes P_{i,N}^{-1}, \) \( V_{i,t} = \Sigma \otimes \left( \frac{1}{N} I_{N \times (N-n)} \right) \), \( P_{i,N}^{-1} I_n = \frac{N}{N-n} I_n \) and

\[
I_n' P_{i,N}^{-1} I_n = \frac{N}{N-n},
\]

we have the following useful identities, \( \left( G_i'' V_i^{-1} G_i^* \right)^{-1} = \frac{N-n}{N} \left( \Sigma^{-1} u_i \right)^{-1} \) with

\[
u_i = \left( 1 \ 0_{1 \times p} \right) \text{ and } V_i^{-1} V_{i,t} = -\frac{1}{N-n} I_{p+1} \otimes J_{n \times (n-N)},
\]

Subsequently, (A.1) can be simplified as

follows,

\[
\dot{w} = \left\{ -\frac{1}{N-n} I_{p+1} \otimes J_{n \times (n-N)} + \frac{1}{n} \left( \Sigma^{-1} u_i \otimes \left( P_{i,N}^{-1} I_{N-n} \right) \right) \right\} C_H
\]

\[
= \left\{ \left( \frac{1}{N-n} I_n \otimes I_n \right) - \left( \Sigma^{-1} u_i \otimes I_n \right) \right\} C_H.
\]

Since \( \Sigma^{-1} = \begin{pmatrix} \sigma_y' & \sigma_y' \\ \sigma_y & \Sigma_y \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_y^2 - \sigma_y \Sigma_y & -\sigma_y \Sigma_y \\ -\Sigma_y \sigma_y & \Sigma_y \end{pmatrix}, \)

\[
\left( \Sigma^{-1} u_i \right) = \begin{pmatrix} \sigma_y^2 - \sigma_y \Sigma_y \Sigma_y^{-1} \\ -\Sigma_y \sigma_y \Sigma_y^{-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{\Sigma_y} \sigma_y \\ \frac{1}{\Sigma_y} \Sigma_y \end{pmatrix} = \begin{pmatrix} \beta \\ \Sigma_y \beta \end{pmatrix}
\]

where \( \beta = \Sigma^{-1} \sigma_y \). Therefore, we have

\[
\dot{w} = \left\{ \frac{1}{N-n} I_n \otimes I_n \right\} C_H.
\]
Consequently,

\[ \hat{P} = (C_i + w)' Z_i' \]

\[ = c'_i Y_i + c'_i 1_{X_i} \left\{ \bar{Y}_i - \frac{N}{N-n} \beta' (\bar{X}_i - \mu_i) \right\} \]

\[ = \sum_{i=1}^{n} c_i Y_i + \sum_{i=n+1}^{N} c_i \left\{ \bar{Y}_i - \frac{N}{N-n} \sum_{k}^{c_i} \beta_k (\bar{X}_k - \mu_k) \right\} \]

and

\[ \text{var} \left( \hat{P} \right) = (C_i + w)' V_j (C_i + w) \]

\[ = \left( \sum_{i=1}^{n} c_i ^2 - \frac{1}{n} \left( \sum_{i=1}^{n} c_i \right)^2 \right) \sigma_0^2 + n(1-f) \left( \frac{1}{n} \sum_{i=1}^{n} c_i - \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right)^2 \sigma_0^2 \]

\[ + 2N(1-f) \left( \frac{1}{n} \sum_{i=1}^{n} c_i - \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right) \left( \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right) (1 - \rho_{YX}^2) \sigma_0^2 \]

\[ + N^2 \frac{(1-f)}{n} \left( \frac{1}{N-n} \sum_{i=n+1}^{N} c_i \right)^2 (1 - \rho_{YX}^2) \sigma_0^2, \]

where \( \rho_{YX}^2 = \sigma_{YX}^2 / \sigma_0^2 \) is the multiple correlation coefficient of \( Y \) on \( X \).
We represent the population by an \( N \times 2 \) matrix \((y \ x)\), where \( y \) and \( x \) are \( N \times 1 \) column vectors of the response and auxiliary values, respectively. The population mean and total for the response and auxiliary variates are given by \( \mu_y = \frac{1}{N} \sum_{j=1}^{N} y_j \) and \( \mu_x = \frac{1}{N} \sum_{j=1}^{N} x_j \) and \( T_y = \sum_{j=1}^{N} y_j \) and \( T_x = \sum_{j=1}^{N} x_j \). We define \( \sigma_y^2 = \frac{1}{N-1} \sum_{j=1}^{N} (y_j - \bar{y})^2 \), \( \sigma_x^2 = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \bar{x})^2 \) and \( \sigma_{xy} = \frac{1}{N-1} \sum_{j=1}^{N} (y_j - \bar{y})(x_j - \bar{x}) \), and summarize the variance and covariance of the two variates by \( \frac{N-1}{N} \Sigma = \frac{N-1}{N} \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix} \).

The values of the permuted units

where \( U_i = (U_{i1} \ U_{i2} \ \cdots \ U_{iN})' \), represent a matrix of indicator ran

\[
E(X) = \mu_x 1_N \quad \text{and} \quad \text{var}(X) = \sigma_x^2 P_N . \]

Using \( \begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} \), we can show that

\[
E \begin{pmatrix} Y \\ X \end{pmatrix} = G \mu , \quad \text{where} \quad G = \begin{pmatrix} 1_N & 0 \\ 0 & 1_N \end{pmatrix} , \quad \mu = (\mu_y \ \mu_x)' , \quad \text{and that} \quad \text{cov} \begin{pmatrix} Y \\ X \end{pmatrix} = \Sigma \otimes P_N .
\]

For clarity, let \( Z = (Y' \ X')' \), the above model can be rewritten in a compact form, \( Z = G \mu + E \).
To incorporate this constraint in the model, we transform $X$ into $X^*$ by subtracting $\mu_x$ from each element of $X$, i.e., $X_j^* = X_j - \mu_x$ or equivalently $X^* = P_N X$. Since $P_N$ is idempotent, the column rank of $X^*$ is the same as that of $X$ and

$$\text{cov}(X^*) = \text{cov}(X).$$

Let $R = \begin{pmatrix} I_N & 0 \\ 0 & P_N \end{pmatrix}$, and define $Z^* = RZ$, $RG = \begin{pmatrix} G^* & 0_{N \times 1} \end{pmatrix}$. The reparameterized model that incorporates the constraint is given by

$$\begin{pmatrix} Y \\ X^* \end{pmatrix} = \begin{pmatrix} 1_{N \times 1} \\ 0_{N \times 1} \end{pmatrix} \mu_y + E^*$$

or in a compact form,

$$Z^* = G^* \mu_y + E^*, \quad (2)$$

where $G^* = \begin{pmatrix} 1_N' & 0_{N \times 1} \end{pmatrix}$, and $E^* = RE$. The partitioned into the sampled $(Z_{ij})$ and remaining $(Z_{il})$ portions through