Building Inference Under Two Stage Cluster Sampling:
A Simple framework useful for Experimental Designs and Longitudinal Studies.

Edward J. Stanek III
Department of Biostatistics and Epidemiology
University of Massachusetts at Amherst, Ma.

Statistical inference can be difficult to describe in a coherent manner when the problem requires discussion of sampling, randomization, prior distributions, and includes fixed and random effects. We present a population framework where two stage cluster sampling gives rise to a random permutation superpopulation. Parameters in the population are linked to models for the superpopulation clarifying inference. We use this framework to address inference in simple hierarchical models, mixed models, and factorial experimental design models. We illustrate the advantage in this approach by using it to help distinguish terms like bias, estimators, predictors, latent values, and random effects.
Building Inference Under Two Stage Cluster Sampling:

Edward J. Stanek III
Department of Biostatistics and Epidemiology
University of Massachusetts at Amherst, Ma.

1. Introduction

Dietary educational interventions delivered by physicians may be an effective way to lower fat intake and positively alter cholesterol. For example, the WATCH Study (1996) in Worcester, Massachusetts was a randomized study designed to lower cholesterol levels in patients of physicians via an educational physician provided dietary intervention. The study had a clustered randomized design. A random sample of physicians in an HMO was randomly assigned to one of three intervention groups, and trained in different interventions to alter their patient’s diet, and in turn, their patient’s cholesterol. A random sample of patients was selected from each selected physician’s practice, with the patient’s cholesterol measured before and after the intervention. One natural question that arose upon completion of the study was how effective were the individual physicians in altering patient’s diets, and lowering their patient’s cholesterol. This question motivated in turn the question: How should one estimate the average cholesterol change for the patients of each physician? Providing a coherent answer to this question is the objective of this presentation.

A simple estimate of the change in cholesterol of physician’s patients is the average change in cholesterol of patients in the physician’s sample. This estimate is intuitive, and broadly accepted by the medical community. At the same time, it is well known that the simple average may be unreliable if only a few patients were sampled for a physician. An alternative estimator, the best linear unbiased predictor (BLUP), is often recommended by statisticians (Robinson, (1991), Harville and Carriquiry, (1992)). This estimator arises from a simple mixed model as:

\[ Y = X\alpha + Z\beta + E \]
where fixed effects (such as treatments) are included in the parameters $\alpha$, while the random effects (for physicians) are contained in $D$. The estimated realized random effect is given by the (BLUP) given by $\bar{y} + k (\bar{y} - \bar{y})$. The factor $k$ in the estimator is a shrinkage constant that depends on the variability of cholesterol change between patients of a physician relative to the variability in average cholesterol change between physicians.

The BLUP estimate has the well known property of being an estimator that has the smallest average MSE among a set of estimators whose average bias is zero. While these properties are appealing, a close reading of the properties illustrates that they are not as impressive as they may appear on first sight. For a realized physician, the BLUP estimate is not unbiased. Furthermore, as illustrated by Stanek and O’Hearn (1998), for a realized physician, the BLUP estimate may not have smaller MSE than the simple sample mean. While better in many ways (see Stanek, Well, and Ockene, (1999) and references therein), there are not always better. This motivates a closer look at inference in the mixed model.

### 1.1 Mixed Models

Mixed models differ from fixed effect linear models in statistics since some parameters are stochastic. This may seem a contradiction in terms, since parameters are usually defined to be deterministic, not stochastic. The stochastic parameters arise from sampling fixed parameters. An understanding of the model, and the ‘random effects’, can be gained by developing the model in a stepwise manner. Such a development highlights the role of sampling. We briefly review this development to define notation and set the context.

Consider a setting where there are $h = 1, \ldots, H$ physician practices (PSUs), each with $t = 1, \ldots, N_h$ (distinct) patients (SSUs). Let us represent the serum cholesterol level (response) of patient $t$ in physician $h$’s practice by $\mu_{ht}$. The parameter $\mu_{ht}$ is a fixed constant, but one
that may differ between patients in different physician practices. We define the average cholesterol for physician $h$’s patients as $\mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \mu_{hh}$. Our goal is to estimate the average cholesterol level of patients in each physician’s practice via data from a two stage simple random sample of $n_h$ patients from a simple random sample of $I$ physician practices. For simplicity, we assume equal sample sizes of patients for each practice ($m = n_h$), and equal practice sizes ($M = N_h$) for all $h = 1, ..., H$. Let response for the $k$th selected patient of physician $h$ be given by $Y_{hk} = \mu + \delta_h + E^*_hk$, where $\mu_h = \mu + \delta_h$. In this model, $E^*_hk$, and hence $Y_{hk}$, is a random variable since it arises from the $k$th selected patient. The response depends on which patient is realized on the $k$th selection in the sample.

Next, consider a model for the $i$th selected physician given by

$$\sum_{h=1}^{H} U_{ih} Y_{hk} = \mu + \sum_{h=1}^{H} U_{ih} \delta_h + \sum_{h=1}^{H} U_{ih} E^*_hk.$$  

$$Y_{ik} = \mu + D_i + E^*_ik.$$  

In this model, we introduce the indicator sampling random variables $U_{ih}$ that have a value of one when the $i$th selected physician is physician $h$, and zero otherwise. The random variable $D_i = \sum_{h=1}^{H} U_{ih} \delta_h$ is a random effect; the randomness clearly arises from the sampling. Note that these random variables are linear combinations of the underlying sampling indicator random variables.
Let a vector of indicator random variables be given by \( S_i = \left( U_{i1}, U_{i2}, \ldots, U_{iH} \right) \), and define the vectors \( Y_k = \left( Y_{1k}, Y_{2k}, \ldots, Y_{Hk} \right) \), \( \delta' = (\delta_1, \delta_2, \ldots, \delta_H) \) and 

\[
E'_k = \left( E_{1k}', E_{2k}', \ldots, E_{Hk}' \right). \]

Then the model

\[
Y_{ik} = \mu + D_i + E_{ik} \\
S_i' Y_k = \mu + S_i' \delta + S_i' E_k'
\]

can be summarized as

\[
SY = X\mu + S\delta \otimes 1_m' + SE'
\]

or

\[
Y = X\mu + ZD + D
\]

where \( S' = (S_1, S_2, \ldots, S_I) \), \( Y = (Y_1, Y_2, \ldots, Y_m) \), \( E^* = (E_{1k}', E_{2k}', \ldots, E_{Hk}') \), 

\[
X = 1_{It} \otimes 1_m', \quad 1_m \text{ is an } m \times 1 \text{ vector of ones, } Y_m = \text{vec} \left[ \left( SY \right)' \right], \quad X = 1_{ml}, \quad Z = I_1 \otimes 1_m, \quad \text{and} \]

\[
D = \text{vec} \left[ \left( SE' \right)' \right]. \]

The model equation makes explicit the role of the sample indicator random variables in the mixed model. In this model, the error, \( D \), is a product of random variables.

### 1.2 Comparison of the Sample Mean and BLUP estimator for a Realized Effect

Of particular interest is a comparison of estimators of the parameter for a realized physician’s practice. Two estimators, the sample mean and the BLUP estimator, were compared relative to the MSE by Stanek and O’Hearn (1998). This comparison was made in a conditional framework (conditional on a realized physician, ie. \( S_i = s_i \)). The comparisons indicate that for physicians whose mean response is close to the population mean, the BLUP
estimator will have smaller MSE than the simple mean estimate. However, for physicians whose mean response is far from the average response, the simple mean will have smaller MSE.

The conditional framework results in a different sample space for the inference for different realized physicians. Such considerations become cumbersome when one considers questions like “How do we estimate the difference in response between two physicians?” since alternative sample spaces will be relevant. Studying such problems will likely become even more complex when the number of patients (and sample sizes) differ between physician practices, or more general linear combinations of parameters are to be estimated. For these reasons, it is of interest to explore other frameworks for the problem that may result in potentially different estimators, and more straightforward development of inference.

1.3 Overview and Objectives of This Paper

We present and develop a framework for estimation of linear functions of the population that is useful in two stage cluster sampling settings. The framework that we propose we call a superpopulation sampling framework. We first describe this superpopulation sampling framework, and contrast it with the superpopulation model framework that is defined in the literature.

Next, we illustrate how this framework can be used to derive estimators and construct inference. The steps in this process closely parallel steps used in constructing inference in the superpopulation sampling literature (Bolfarine and Zacks (1992)), where the objective is prediction of unobserved superpopulation random variables. An added step in developing inference is the direct connection of parameters in the population with parameters in the
model for the superpopulation. These relationships imply constraints on the design matrix in the model for the superpopulation.

In an effort to keep the focus on the principal ideas, we first develop the estimators in the context of a simple problem where there is single stage simple random sampling without replacement. Next, we repeat the development with two stage sampling, where the second stage is considered to be sampling from an infinite population (or response error). We follow this with two stage finite population sampling, with equal numbers of SSUs per PSU. The objective is development of an optimal estimator of primary sample unit means based on a sample. This development establishes the basic strategy for inference, the limitations of this strategy, and establishes basic notation.

We conclude with a discussion of how this framework and strategy can be extended to two stage finite population sampling settings. The first problem we consider is for two stage cluster sampling (using simple random sampling without replacement at both stages). We also consider how the inferential framework can be applied to a simple factorial experimental design. We conclude by summarizing new ideas that are a consequence of the earlier developments, and outline the future direction of this research.

2. Superpopulations in the Literature

Superpopulations have been introduced in survey sampling to facilitate modelling survey data (as in Hartley and Sielken (1975), Cassel, Sarndal, and Wretman (1977), Bolfarine and Zacks (1992), Nandram and Sedransk (1993), Tam (1995), Cocchi and Mouchart (1996), Arora, Lahiri, and Mukherjee (1997), and Nandram (1999)). Cassel, Sarndal and Wretman (1977) describe several ways in which superpopulation models may be
conceptualized. In all cases, a distinction is made between the population (a fixed set of constants indexed by $h = 1, ..., H$) and the superpopulation (a similar set of random variables). One common conceptualization is that the finite population is a realization of a set of random variables that make up the superpopulation (see for example, Hartley and Sielken (1975), Royall (1992), and Tam (1995)). A model is hypothesized for the superpopulation, and the goal is to predict functions of the realized population values based on observations on a sample of the population. Using this concept, the superpopulation may be thought of as a larger universe (possibly infinite) from which the population is drawn. The population is represented as a non-stochastic vector of fixed constants (ie. $y_1, y_2, \cdots, y_H$), whereas the superpopulation is a similar vector of random variables (ie. $Y_1, Y_2, \cdots$).

A simple superpopulation model is the random permutation superpopulation model (see Cassel, Sarndal, and Wretman, p87). In this model, the superpopulation consists of random variables corresponding to permutations of units in the finite population, where permutations occur with equal probability. The population is particular permutation. Parameters in the population consist of the population values, and functions of these parameters. Superpopulation parameters are defined as the expected value (over the superpopulation) of functions of superpopulation random variables (such as a regression parameter).

The superpopulation model approach distinguishes the randomness arising from sampling from the randomness due to the modeling. The object of the analysis is usually prediction of a function of parameters in the population (such as $\bar{y}$) based on their partial observation. Estimators are chosen that minimize a superpopulation model based expected value of the design based MSE (see p92 and section 5.5 of Cassel, Sarndal, and Wretman.
A concern in superpopulation model based inference is design bias. There is an imaginary quality of the traditional notion of a superpopulation (see Hartley and Sielken, (1975)), which some authors argue may be an advantage (see Barnard (1973)).

3. A Superpopulation Framework for Inference of Linear Functions: Simple Random Sampling

3.1 Definition of the Population and the Superpopulation

3.1.1 The Population

The finite population consists of \( h = 1, \ldots, H \) primary sampling units (PSUs) with non-stochastic values given by \( \mathbf{y} = (y_1, y_2, \ldots, y_H) \).

3.1.2 The Superpopulation

We define a superpopulation that represents permutations of the PSUs the finite population. We label the order of the units in a permutation with the subscript \( i = 1, \ldots, H \). In general, the unit assigned to the first position in the permutation \( i = 1 \) could be any one of the \( H \) units. We represent the possible assignment of a unit to the first position via \( H \) random variables \( Y_{11}, Y_{12}, \ldots, Y_{1H} \), where the value of the random variable \( Y_{1h} \) has a value of \( y_h \) if unit \( h \) is assigned to the first position, and a value of 0 if unit \( h \) is not assigned to the first position. More generally, we define a set of random variables that indicate whether a particular unit is assigned to a particular position as \( U_{ih} \), where the value of \( U_{ih} \) is equal to 1 if unit \( h \) is assigned to position \( i \), and zero otherwise.

We summarize these random variables in the following expression:
In this expression, the vector of random variables \((Y_1, Y_2, \ldots, Y_H)\)' represents a random permutation of the PSUs in the population. Notice that each random variable in this vector consists of sums of \(H\) random variables.

We define the superpopulation to be a vector of the basic \(H^2\) random variables of the form \(Y_{ih} = U_{ih} \cdot y_h\). To define the vector of random variables compactly, we introduce notation for a diagonal matrix. Let us represent a diagonal matrix with diagonal elements equal to the elements of the vector \(\mathbf{a}\) as \(\mathbf{D}_a\). Then we defined the superpopulation as the \(H^2 \times 1\) vector of random variables given by \(\mathbf{Y} = \mathbf{D}_{(y_{1h})} \cdot \text{vec}(\mathbf{U})\).

The superpopulation vector contains the basic random variables that arise from sampling the finite population. The superpopulation is the full set of \(H^2\) random variables that arise from the \(i = 1, \ldots, H\) selections of population PSUs in a permutation, where

\[
\mathbf{Y}' = \begin{pmatrix} Y_1' & \cdots & Y_h' & \cdots & Y_H' \end{pmatrix}
\quad \text{and} \quad
\mathbf{Y}_h' = \begin{pmatrix} Y_{1h} & \cdots & Y_{ih} & \cdots & Y_{Hh} \end{pmatrix}
\]

such that

\[
\mathbf{Y}' = \left( \begin{pmatrix} Y_{11} & Y_{21} & \cdots & Y_{H1} \\ Y_{12} & Y_{22} & \cdots & Y_{H2} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1H} & Y_{2H} & \cdots & Y_{HH} \end{pmatrix} \right)
\quad \text{and} \quad
\mathbf{Y}_h' = \begin{pmatrix} Y_{1h} & Y_{2h} & \cdots & Y_{Hh} \end{pmatrix}
\]

where \(Y_{ih} = U_{ih} \cdot y_h\). The random variables \(U_{ih}\) are indicator sampling random variables that have a value of one when the \(i^{th}\) PSU in the permutation corresponds to PSU \(h\), and zero
otherwise. These basic random variables give rise to random variables that represent an individual PSU in a permutation, such as \( Y_i = \sum_{h=1}^{H} U_{ih} y_h \).

The superpopulation captures the basic indicator random variables generated by a simple random without replacement sample in the spirit of Neyman’s (1923) man-made randomness. The superpopulation is artificial (but not imaginary), in the sense that it spans a larger dimensional space (ie. \((H-1)^2\)) than the population. However, the superpopulation directly accounts for the random variables created by the random sampling, while not reducing the dimension of the basic sampling random variables. It differs from the random permutation superpopulation model defined by Cassel, Sarndal and Wretman (1977) since it contains \( H^2 \) random variables, as opposed to \( H \) random variables.

### 3.1.3 The relationship between the Superpopulation and the Population

There is a structural relationship between the superpopulation and the population. The random superpopulation vector can be projected onto the population using the projection \( C' = I_H \otimes 1_H \) such that

\[
C'Y = C'(y \otimes 1_H)_{\text{vec}}(U)
= \left( I_H \otimes 1_H \right)'D_{(y \otimes 1_H)}\text{vec}(U) \\
= \left( D_y \otimes 1_H \right)\text{vec}(U) \\
= \text{vec} \left( 1_H y' UD_y \right) \\
= \text{vec} \left( 1_H y' D_y \right) \\
= y
\]
Elements of \( C'Y \) have the form \( \sum_{i=1}^{H} U_{ih} y_h = y_h \left( \sum_{i=1}^{H} U_{ih} \right) \). Since the superpopulation is a permutation of the PSUs, PSU \( h \) will occupy only one position in a given permutation of PSUs (implying that \( \sum_{i=1}^{H} U_{ih} = 1 \)), and hence that \( \sum_{i=1}^{H} U_{ih} y_h = y_h \) for all permutations. This projection is unique.

Although there is only one projection that will reduce the random variables in the superpopulation to the population values (for any realization of the superpopulation), there are other projections of the superpopulation where the expected value of the projection (over realizations of the superpopulation) will equal the population values. As an example, consider the projection \( \tilde{C}'Y \) where \( \tilde{C}' = \bigoplus_{h=1}^{H} w_h \) where \( w_h = (w_{1h}, w_{2h}, \ldots, w_{Hh}) \). Then elements of \( \tilde{C}'Y \) are given by \( y_h \sum_{i=1}^{H} w_{ih} U_{ih} \). If for all \( h = 1, \ldots, H \),

\[
E_{\pi}\left( \sum_{i=1}^{H} w_{ih} U_{ih} \right) = \sum_{i=1}^{H} w_{ih} E_{\pi}(U_{ih}) = 1,
\]

then the expected value of the projected superpopulation will equal the population.

**3.2 Population Parameters and Models for the Superpopulation**

The objective is estimation of parameters that can be expressed as linear functions of the values in the finite population. The functions correspond to individual PSU means. We first define these parameters in the population. Next, we define a model for the superpopulation and relate the superpopulation parameters to the population parameters. In addition, we specify any additional unbiased constraints and assumptions. Finally, we
develop expressions for the expected value and variance of the superpopulation random variables.

3.2.1 Population Parameters

The population values (ie. \( y' = (y_1, y_2, \ldots, y_H) \)) are the population parameters, \( \mu' = (\mu_1, \mu_2, \ldots, \mu_H) \), and thus \( y = \mu \). We can express the parameters as a linear function of the population values, \( \mu = C' y \) where \( C^* = I_H \), or the values as a function of the parameters plus a residual, such that \( y = X' \mu + e \), where \( X^* = I_H \) and \( e = 0 \).

3.2.2 A Model for the Superpopulation

We define a model for the superpopulation as \( Y_{ih} = x_{ih} \beta_h + E_{ih} \) for \( i = 1, \ldots, H \) and \( h = 1, \ldots, H \), where \( x_{ih} \) is a known constant, and summarize this model as \( Y = X\beta + E \).

The superpopulation model contains a deterministic and random component. Without additional assumptions, there is no explicit connection between the parameters in the superpopulation \( \beta \), and the population parameters \( \mu \).

3.2.3 The Relationship between Superpopulation Parameters and Population Parameters

We relate the superpopulation parameters to the population parameters by equating the non-stochastic portion of the superpopulation model when projected onto the population, to the non-stochastic portion of the population model containing the parameters of interest.

Using the projection, \( C' Y = y = C' X\beta + C' E \), the non-stochastic portion of the superpopulation model when projected onto the population is given by \( C' X\beta \). In the
population model, \( y = X\beta + e \), the non-stochastic portion containing the parameters of interest is given by \( X\beta \). We define the superpopulation parameters such that \( C'X\beta = X\mu \).

Without loss of generality, we further require that \( \beta = \mu \). Even with this assumption, the design matrix \( X \) is not uniquely defined. One design matrix that will satisfy this constraint is

\[ X = I_H \otimes \frac{1_H}{H}. \]

We use this design matrix in the superpopulation model.

Since \( \beta = \mu \), \( \mu = C'y \), and \( C'Y = y \), we can express the superpopulation parameters as a linear function of the superpopulation. Thus, \( \beta = C'C'Y \).

### 3.3 Expected Value and Variance of the Superpopulation

#### 3.3.1 Expected Value of Superpopulation Vector

We construct expressions for the expected value and variance of the superpopulation using the subscript \( \xi \) to indicate expectation over sampling. First note that elements of \( Y \) are given by \( Y_{ih} = U_{ih}y_{ih} \). Also, under the random permutation model,

\[ E_{\xi}(U_{ih}) = \frac{1}{H} \text{ for } i = 1, \ldots, H \text{ and } h = 1, \ldots, H, \] so that \( E_{\xi}(Y_{ih}) = E_{\xi}(U_{ih})y_{ih} = \frac{y_{ih}}{H} \). As a result,

\[ E_{\xi}(Y_{ih}) = E_{\xi}(Y_{ih} \cdots Y_{ih}) = \frac{y_{ih}}{H}1_H' \], and since \( Y' = \begin{pmatrix} Y_1' \cdots Y_H' \end{pmatrix} \),

\[ E_{\xi}(Y) = y \otimes \frac{1_H}{H}. \]

#### 3.3.2 Variance of the Superpopulation Vector

We evaluate an expression for the variance of the superpopulation next. First, note that

\[ \text{var}_{\xi}(Y_{ih}) = E_{\xi}[Y_{ih}Y_{ih}'] - E_{\xi}[Y_{ih}]E_{\xi}[Y_{ih}']. \]

We have evaluated \( E_{\xi}(Y_{ih}) = y_{ih} \otimes \frac{1_H}{H} \), and
hence \( \mathbb{E}_{\xi_1}[Y_h] \mathbb{E}_{\xi_1}[Y_h'] = y_h^2 \odot \frac{J_H}{H^2} \). Now

\[
\mathbb{E}_{\xi_1}(Y_h Y_h') = y_h^2 \mathbb{E}_{\xi_1}\left[ \begin{pmatrix} U_{1h} \\ U_{2h} \\ \vdots \\ U_{Hh} \end{pmatrix} \right] = y_h^2 \mathbb{E}_{\xi_1}\left[ \begin{pmatrix} U_{1h} \\ U_{2h} \\ \vdots \\ U_{Hh} \end{pmatrix} \right].
\]

This expected value is made up of terms like \( \mathbb{E}_{\xi_1}(U_{ih}^2) = \frac{1}{H} \) for \( i = 1, \ldots, H \) and terms like

\[
\mathbb{E}_{\xi_1}(U_{ih} U_{i'h}) = 0 \text{ for } i = 1, \ldots, H \text{ and } i^* (\neq i) = 1, \ldots, H.
\]

Thus,

\[
\mathbb{E}_{\xi_1}(Y_h Y_{h^*}) = \frac{I_H}{H} \text{ and } \operatorname{var}_{\xi_1}(Y_h') = \frac{y_h^2}{H} \left( I_H - \frac{J_H}{H} \right).
\]

We also need to evaluate \( \operatorname{cov}_{\xi_1}(Y_h Y_{h^*}) = \mathbb{E}_{\xi_1}(Y_h Y_{h^*}) - \mathbb{E}_{\xi_1}(Y_h) \mathbb{E}_{\xi_1}(Y_{h^*}) \) for \( h \neq h^* \).

Now \( \mathbb{E}_{\xi_1}(Y_h Y_{h^*}) = y_h y_{h^*} \mathbb{E}_{\xi_1}\left[ \begin{pmatrix} U_{1h} \\ U_{2h} \\ \vdots \\ U_{Hh} \end{pmatrix} \right] \). When \( i = i^* \),

\[
\mathbb{E}_{\xi_1}(U_{ih} U_{i'h}) = 0 \text{ since the same selection cannot result in two PSUs, while when } i \neq i^* ,
\]

\[
\mathbb{E}_{\xi_1}(U_{ih} U_{i'h}) = \frac{1}{H (H-1)}. \text{ As a result, } \mathbb{E}_{\xi_1}(Y_h Y_{h^*}) = \frac{y_h y_{h^*}}{H (H-1)} \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix}
\]

expressed as \( \mathbb{E}_{\xi_1}(Y_h Y_{h^*}) = \frac{y_h y_{h^*}}{H (H-1)} (J_H - I_H) = \frac{-y_h y_{h^*}}{H (H-1)} \left( I_H - \frac{J_H}{H} \right) + \frac{y_h y_{h^*}}{H^2} J_H \). Then

\[
\operatorname{cov}_{\xi_1}(Y_h Y_{h^*}) = \frac{-y_h y_{h^*}}{H (H-1)} \left( I_H - \frac{J_H}{H} \right) + \frac{y_h y_{h^*}}{H^2} J_H - \frac{y_h y_{h^*}}{H (H-1)} \left( I_H - \frac{J_H}{H} \right).
\]
Combining terms,

\[
\text{var}_y(Y) = \begin{pmatrix}
\frac{y_1^2}{H} (I_H - \frac{J_H}{H}) & \frac{-y_1 y_2}{H} (I_H - \frac{J_H}{H}) & \ldots & \frac{-y_1 y_H}{H} (I_H - \frac{J_H}{H}) \\
\frac{-y_2 y_1}{H} (I_H - \frac{J_H}{H}) & \frac{y_2^2}{H} (I_H - \frac{J_H}{H}) & \ldots & \frac{-y_2 y_H}{H} (I_H - \frac{J_H}{H}) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{-y_H y_1}{H} (I_H - \frac{J_H}{H}) & \frac{-y_H y_2}{H} (I_H - \frac{J_H}{H}) & \ldots & \frac{y_H^2}{H} (I_H - \frac{J_H}{H})
\end{pmatrix}
\]

or

\[
\text{var}_y(Y) = \frac{1}{H(H-1)} \begin{pmatrix}
(H-1)y_1^2 & -y_1 y_2 & \ldots & -y_1 y_H \\
-y_2 y_1 & (H-1)y_2^2 & \ldots & -y_2 y_H \\
\vdots & \vdots & \ddots & \vdots \\
-y_H y_1 & -y_H y_2 & \ldots & (H-1)y_H^2
\end{pmatrix} \otimes \left( I_H - \frac{J_H}{H} \right). \]

Let

\[
D_y = \begin{pmatrix}
y_1 & 0 & \ldots & 0 \\
0 & y_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & y_H
\end{pmatrix}
\]

Then

\[
\begin{pmatrix}
(H-1)y_1^2 & -y_1 y_2 & \ldots & -y_1 y_H \\
-y_2 y_1 & (H-1)y_2^2 & \ldots & -y_2 y_H \\
\vdots & \vdots & \ddots & \vdots \\
-y_H y_1 & -y_H y_2 & \ldots & (H-1)y_H^2
\end{pmatrix} = HD_y \left( I_H - \frac{J_H}{H} \right) D_y,
\]

and hence

\[
\text{var}_y(Y) = \frac{1}{(H-1)} D_y \left( I_H - \frac{J_H}{H} \right) D_y \otimes \left( I_H - \frac{J_H}{H} \right).
\]

### 3.4 Sampling, Re-arranging, and Partitioning

We specify a re-arrangement of the superpopulation vector so that it can be partitioned into the sampled and remaining portions. The sampled portion of the superpopulation we consider to be realized and non-stochastic. We focus attention on
estimation of the remainder of the population. We specify a criteria for estimation, and then derive estimators (or estimating equations) that optimize this criteria.

3.4.1 Re-arranging and Partitioning the Superpopulation

Sampling corresponds to selection of a simple random without replacement sample of \( I \) PSU’s. We re-arrange terms in the superpopulation model into a sampled and remainder vector by pre-multiplying \( Y \) by an \( H^2 \times H^2 \) permutation matrix

\[
K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = \begin{pmatrix} I_H \otimes \left( I_I \begin{pmatrix} 0 \\ I_{s(H-I)} \end{pmatrix} \right) \\ I_H \otimes \left( \begin{pmatrix} 0 \\ (H-I)_{I} \end{pmatrix} \right) I_H^{-1} \end{pmatrix},
\]

and partition the results such that

\[
KY = \begin{pmatrix} K_{11} Y \\ K_{12} Y \end{pmatrix} = \begin{pmatrix} Y_s \\ Y_r \end{pmatrix}.
\]

Note that since \( K \) is a permutation matrix, \( K^TK = I \). The re-arranged superpopulation model is given by

\[
KY = KX\beta + KE
\]

which we represent as a partitioned model such

\[
\begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix} \beta + \begin{pmatrix} E_s \\ E_r \end{pmatrix}.
\]

After selecting the sample, the vector \( Y_s \) will be realized. In this example, since \( X = I_H \otimes \frac{I_H}{H} \), \( X_s = I_H \otimes \frac{I_s}{H} \) and \( X_r = I_H \otimes \frac{I_{H-I}}{H} \).

We partition the variance of the superpopulation in a similar manner. After re-arranging random variables into the sample and remaining portions,

\[
\text{var}_{\eta} \left[ KY \right] = \text{var}_{\eta} \left[ \begin{pmatrix} Y_s \\ Y_r \end{pmatrix} \right],
\]

which we represent as

\[
\text{var}_{\eta} \left[ \begin{pmatrix} Y_s \\ Y_r \end{pmatrix} \right] = \begin{pmatrix} V_s \\ V_r \end{pmatrix}.
\]

3.4.2 Partitioning the Parameters into functions of the Sample And Remaining Superpopulation Vectors.
We express the parameters $\beta$ as a linear combination of the sampled and remaining superpopulation vectors. To do so, recall that $\beta = C\beta Y$. Then, introducing the permutation matrix,

$$\beta = \left(C'K\right)KY = \left(C'K_{11} \quad C'K_{12}\right)\begin{pmatrix}K_{11}Y \\ K_{12}Y\end{pmatrix},$$

$$= \left(L_s' \quad L_r'\right)\begin{pmatrix}Y_s' \\ Y_r'\end{pmatrix} = L_s'Y_s + L_r'Y_r,$$

where $L_s' = C'K_{11}$ and $L_r' = C'K_{12}$. These expressions simplify to $L_s' = I_H \otimes I_i'$ and $L_r' = I_H \otimes I_{H-i}$.

### 3.5 Estimation

Our goal is to estimate $\beta$ based on the realized sample. Since $\beta = L_s'Y_s + L_r'Y_r$ and $Y_s$ is realized, the target of estimation is $L_r'Y_r$. The values of $L_r'Y_r$ are not observed, and hence the problem is commonly described as prediction of $L_r'Y_r$. We use this terminology.

#### 3.5.1. Properties of Predictors

We consider predictors $L_{rs'}$ that are:

1. linear in the data (of the form $L_{rs'}Y_s$) and

2. are unbiased (such that $E_{\psi_i}\left(L_{rs'}Y_s\right) = E_{\psi_i}\left(L_r'Y_r\right)$).
The unbiased constraint can be expressed as a constraint on the predictors \( L_{rs} \). First, note that since \( X = H \otimes_1 \frac{1_H}{H} \) and \( \beta = \mu = y \), \( X\beta = \left( \frac{1_H}{H} \right) y = y \otimes_1 \frac{1_H}{H} \). Also, \( E_{\xi_i}(Y) = y \otimes \frac{1_H}{H} \).

Now the superpopulation model is given as \( Y = X\beta + E \), and hence \( E_{\xi_i}(Y) = X\beta \) and \( E_{\xi_i}(E) = 0 \). As a result, \( E_{\xi_i}(KY) = E_{\xi_i}\left[ \begin{array}{c} Y_s \\ Y_r \end{array} \right] = \left[ \begin{array}{c} X_s \beta \\ X_r \beta \end{array} \right], \) and the unbiased constraint simplifies to \( L_{rs}' X_s \beta = L_r' X_r \beta \). This implies that that the estimator must satisfy the constraint that \( L_{rs}' X_s = L_r' X_r \).

3.5.2. Optimization Criteria

We consider the best estimator to be one that satisfies the constraint that \( L_{rs}' X_s = L_r' X_r \), and minimizes the generalized mean squared error (see Bolfarine and Zacks, 1992, p17) given by

\[
\text{GMSE} = 1_H' E\left[ \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)' \right] 1_H
\]

\[
= 1_H' \left( L_{rs}' \mid -L_r' \right) \text{var} \left[ \begin{array}{c} Y_s \\ Y_r \end{array} \right] \left( L_{rs}' \mid -L_r' \right)' 1_H
\]

\[
= 1_H' \left[ L_{rs}' V_s L_{rs} - L_{rs}' V_r L_r - L_r' V_s L_{rs} + L_r' V_r L_r \right] 1_H
\]

The resulting estimator we denote by \( L_{rs}' \).
The product $\mathbf{I}_H \mathbf{L}_{rs} \mathbf{'} = \sum_{h=1}^{H} \mathbf{L}_{rs:h}$ is a vector of dimension $1 \times HI$. Let us represent elements of $\mathbf{L}_{rs} = \begin{pmatrix} \mathbf{L}_{rs:1} \mathbf{'} & \mathbf{L}_{rs:2} \mathbf{'} & \cdots & \mathbf{L}_{rs:H} \mathbf{'} \end{pmatrix}$ where $\mathbf{L}_{rs:h}$ are row vectors of dimension $1 \times I$. Then we can represent $\mathbf{I}_H \mathbf{L}_{rs} \mathbf{'} = \left( \sum_{h=1}^{H} \mathbf{L}_{rs:h} \right)^{'} = \sum_{h=1}^{H} \mathbf{L}_{rs:h} \mathbf{'}$. This product will have $HI$ elements. Let us represent $\mathbf{L}_{rs:h} = \begin{pmatrix} \mathbf{L}_{rs:h,1} \mathbf{'} & \mathbf{L}_{rs:h,2} \mathbf{'} & \cdots & \mathbf{L}_{rs:h,I} \mathbf{'} \end{pmatrix}$ more generally $\mathbf{L}_{rs:h} = \begin{pmatrix} \mathbf{L}_{rs:h,1} \mathbf{'} & \mathbf{L}_{rs:h,2} \mathbf{'} & \cdots & \mathbf{L}_{rs:h,I} \mathbf{'} \end{pmatrix} = \sum_{h=1}^{H} \mathbf{L}_{rs:h} \mathbf{'}$. Then the expression for the generalized MSE is a function of this $1 \times HI$ vector. We seek values of the elements of $\mathbf{L}_{rs:h}$ that will minimize the GMSE.

The expression for the GMSE must be minimized with respect to $\mathbf{L}_{rs:h}$ subject to the constraint that $\mathbf{L}_{rs:h} \mathbf{'} \mathbf{X}_s = \mathbf{L}_{rs:h} \mathbf{'} \mathbf{X}_r$. This constraint implies a constraint on the elements in $\mathbf{L}_{rs:h} \mathbf{'}$. Since $\mathbf{L}_{rs:h} \mathbf{'} = \mathbf{I}_H \mathbf{L}_{rs:h} \mathbf{'}$, the corresponding constraint in terms of these elements is given by

$\mathbf{I}_H \mathbf{L}_{rs:h} \mathbf{'} \mathbf{X}_s = \mathbf{I}_H \mathbf{L}_{rs:h} \mathbf{'} \mathbf{X}_r$, or $\mathbf{L}_{rs:h} \mathbf{'} \mathbf{X}_s = \mathbf{I}_H \mathbf{L}_{rs:h} \mathbf{'} \mathbf{X}_r$.

### 3.6. Constructing the Estimating Equations
We construct the estimating equations by expressing the GMSE in terms of

such that

\[
\text{GMSE} = \left[ L_{rs} \cdot V_{zs} L_{rs} L_{rs} \cdot 1_{H} - L_{rs} \cdot V_{rs} L_{rs} L_{rs} \cdot 1_{H} - L_{r} \cdot V_{rs} L_{rs} L_{rs} \cdot 1_{H} + 1_{H} \cdot L_{r} \cdot V_{r} L_{r} \cdot 1_{H} \right].
\]

We minimize the GMSE with respect to \( L_{rs} \) subject to the constraint that

\[
L_{rs} \cdot X_{s} = 1_{H} \cdot L_{r} \cdot X_{r}
\]

using Lagrangian multipliers. Let us define

\[
X_{s} = \left( X_{s_{1}}, X_{s_{2}}, \ldots, X_{s_{H}} \right), \quad \text{and} \quad X_{r} = \left( X_{r_{1}}, X_{r_{2}}, \ldots, X_{r_{H}} \right)
\]

Then we can express the constraint as the set of \( h = 1, \ldots, H \) equations

\[
L_{rs} \cdot X_{s} = 1_{H} \cdot L_{r} \cdot X_{r} = 0.
\]

Using these equations, we define the Lagrangian that is to be minimized as

\[
\text{LR} = L_{rs} \cdot V_{zs} L_{rs} L_{rs} \cdot 1_{H} - L_{rs} \cdot V_{rs} L_{rs} L_{rs} \cdot 1_{H} - L_{r} \cdot V_{rs} L_{rs} L_{rs} \cdot 1_{H} + 1_{H} \cdot L_{r} \cdot V_{r} L_{r} \cdot 1_{H}
\]

\[
+ \sum_{h=1}^{H} 2\lambda_{h} \left( X_{s_{h}} L_{rs} - X_{r_{h}} L_{r} \right).
\]

We differentiate this equation with respect to \( L_{rs} \) and \( \lambda_{h} \) for \( h = 1, \ldots, H \), and then set the resulting derivatives to zero. We make use of standard matrix differentiation results (Harville, 1997 #1031)(p295) such that

\[
\frac{\partial (x A x) \partial x}{\partial x} = (A + A') x, \quad \frac{\partial (A' x) \partial x}{\partial x} = A
\]

such that

\[
\frac{\partial (A' x)}{\partial x'} = A'.
\]

The derivative with respect to \( L_{rs} \) will result in the equation

\[
\frac{\partial \text{LR}}{\partial L_{rs}} = \frac{\partial \text{LR}}{\partial L_{rs}} + \sum_{h=1}^{H} 2\lambda_{h} \left( \frac{\partial X_{s_{h}} L_{rs}}{\partial x} - \frac{\partial X_{r_{h}} L_{r}}{\partial x} \right).
\]
\[
\frac{\partial LR(h)}{\partial L_{rs}} = (2V_r L_{rs} - 2V_r L_H, 1) + 2 \sum_{h=1}^{H} \lambda_h X_{sh} \quad \text{and the derivative} \quad \frac{\partial LR(h)}{\partial \lambda_h} \quad \text{with respect to} \quad \lambda_h
\]

for \( h = 1, \ldots, H \) will result in the equations \( \frac{\partial LR(h)}{\partial \lambda_h} = 2 \left( \sum_{h=1}^{H} \lambda_h X_{sh} \right) \). Setting these equations equal to zero, the best linear unbiased predictor is the solution to the equations:

\[
(2V_r L_{rs} - 2V_r L_H, 1) + 2 \sum_{h=1}^{H} \lambda_h X_{sh} = 0 \quad \text{and} \quad 2 \left( \sum_{h=1}^{H} \lambda_h X_{sh} \right) = 0 \quad \text{for} \quad h = 1, \ldots, H.
\]

We can express these equations simultaneously as

\[
\begin{pmatrix}
V_r \\
X_{s1}' \\
X_{s2}' \\
\vdots \\
X_{sh}'
\end{pmatrix}
\begin{pmatrix}
X_{s1} \\
X_{s2} \\
\vdots \\
X_{sh}
\end{pmatrix}
\begin{pmatrix}
\cdots \\
\cdots \\
\cdots \\
\cdots
\end{pmatrix}
= 0
\]

or equivalently as

\[
\begin{pmatrix}
V_r \\
X_{s1}' \\
X_{s2}' \\
\vdots \\
X_{sh}'
\end{pmatrix}
\begin{pmatrix}
L_{rs} \\
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_H
\end{pmatrix}
= \begin{pmatrix}
V_r L_H \\
X_{s1}' L_H \\
X_{s2}' L_H \\
\vdots \\
X_{sh}' L_H
\end{pmatrix}
\]

### 3.7. Simplifications of the Estimating Equations

#### 3.7.1 General Solution to the Estimating Equations

We give a general solution to a set of estimating equations of the form

\[
\begin{pmatrix}
T \\
U
\end{pmatrix}
\begin{pmatrix}
M \\
\lambda \lambda
\end{pmatrix}
= \begin{pmatrix}
A \\
B
\end{pmatrix}.
\]

We can solve these equations explicitly. Note from (Harville, 1997...
#1031)(p99), that

\[
\begin{pmatrix}
\mathbf{T} & \mathbf{U}' & 0
\end{pmatrix}
\begin{pmatrix}
\mathbf{T} & \mathbf{U}' & 0
\end{pmatrix}^\top =
\begin{pmatrix}
\mathbf{T}^\top + \mathbf{T}^\top \mathbf{U} \mathbf{U}^\top & \mathbf{T}^\top \mathbf{U} \\
-\mathbf{Q} \mathbf{U}^\top & \mathbf{Q}^\top
\end{pmatrix}
\] where \( \mathbf{Q} = -\mathbf{U} \mathbf{U}^\top \), or equivalently,

\[
\begin{pmatrix}
\mathbf{T} & \mathbf{U}' & 0
\end{pmatrix}
\begin{pmatrix}
\mathbf{T} & \mathbf{U}' & 0
\end{pmatrix}^\top =
\begin{pmatrix}
\mathbf{T}^\top & -\mathbf{U} \mathbf{U}^\top \mathbf{U} \mathbf{U}^\top & \mathbf{T} \mathbf{U} (\mathbf{U} \mathbf{U}^\top)^\top \\
(\mathbf{U} \mathbf{U}^\top)^\top \mathbf{U}^\top & -\mathbf{U} \mathbf{U}^\top
\end{pmatrix}
\]

Then

\[
\begin{pmatrix}
\mathbf{M} \\
\mathbf{x}
\end{pmatrix} = \begin{pmatrix}
\mathbf{T} & \mathbf{U}' & 0
\end{pmatrix} \begin{pmatrix}
\mathbf{A} \\
\mathbf{B}
\end{pmatrix} = \begin{pmatrix}
\mathbf{T}^\top & -\mathbf{T} \mathbf{U} (\mathbf{U} \mathbf{U}^\top)^\top \mathbf{U}^\top & \mathbf{T} \mathbf{U} \mathbf{U}^\top \\
(\mathbf{U} \mathbf{U}^\top)^\top \mathbf{U}^\top & -\mathbf{U} \mathbf{U}^\top
\end{pmatrix} \begin{pmatrix}
\mathbf{A} \\
\mathbf{B}
\end{pmatrix}.
\]

In particular, \( \mathbf{M} = \begin{pmatrix}
\mathbf{T}^\top & -\mathbf{T} \mathbf{U} (\mathbf{U} \mathbf{U}^\top)^\top \mathbf{U}^\top & \mathbf{T} \mathbf{U} \mathbf{U}^\top \\
(\mathbf{U} \mathbf{U}^\top)^\top \mathbf{U}^\top & -\mathbf{U} \mathbf{U}^\top
\end{pmatrix} \mathbf{A} + \mathbf{T} \mathbf{U} (\mathbf{U} \mathbf{U}^\top)^\top \mathbf{B} \). 

We equate terms in this solution to corresponding terms in our problem, where

\[
\mathbf{M} = \mathbf{L}_{rs}, \quad \mathbf{T} = \mathbf{V}_s, \quad \mathbf{U} = \mathbf{X}_s, \quad \mathbf{A} = \mathbf{V}_s \mathbf{L}_r \mathbf{1}_H, \quad \text{and} \quad \mathbf{B} = \mathbf{X}_r \mathbf{L}_r \mathbf{1}_H .
\]

Thus,

\[
\mathbf{L}_{rs} = \left( \mathbf{V}_s^\top - \mathbf{V}_s^\top \mathbf{X}_s \left( \mathbf{X}_s^\top \mathbf{V}_s \mathbf{X}_s \right)^\top \mathbf{X}_s \right)^\top \mathbf{V}_s \mathbf{L}_r \mathbf{1}_H + \mathbf{V}_s^\top \mathbf{X}_s \left( \mathbf{X}_s^\top \mathbf{V}_s \mathbf{X}_s \right)^\top \mathbf{X}_r \mathbf{L}_r \mathbf{1}_H .
\]

Since \( \mathbf{L}_{rs} = \mathbf{L}_{rs} \mathbf{1}_H \), one possible solution to the equations is given by

\[
\mathbf{L}_{rs} = \left( \mathbf{V}_s^\top - \mathbf{V}_s^\top \mathbf{X}_s \left( \mathbf{X}_s^\top \mathbf{V}_s \mathbf{X}_s \right)^\top \mathbf{X}_s \right)^\top \mathbf{V}_s \mathbf{L}_r + \mathbf{V}_s^\top \mathbf{X}_s \left( \mathbf{X}_s^\top \mathbf{V}_s \mathbf{X}_s \right)^\top \mathbf{X}_r \mathbf{L}_r .
\]

### 3.7.2. Explicit Expressions for the Solution

We present expressions for terms needed in the estimating equation to form a solution. For this problem, expressions are given for \( \mathbf{L}_s, \mathbf{L}_r, \mathbf{X}_s, \mathbf{X}_r, \mathbf{V}_s \), and \( \mathbf{V}_s \) in
previous sections. From section 3.4.2, \( \mathbf{L}_s = \mathbf{I}_H \otimes \mathbf{I}_l \) and \( \mathbf{L}_r = \mathbf{I}_H \otimes \mathbf{I}_{H-l} \). Also, from section 3.4.1, \( \mathbf{sH} = \mathbf{I}_H \otimes \mathbf{l} \) and \( \mathbf{rHH} = \mathbf{I}_H \otimes \mathbf{l} \).

Now, \( \mathbf{V}_s = \mathbf{K}_{11} \text{var}_s (\mathbf{Y}) \mathbf{K}_{11} \) and
\[
\text{var}_s (\mathbf{Y}) = \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \otimes \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \text{ while } \mathbf{K}_{11} = \mathbf{I}_H \otimes \left( \mathbf{I}_l \mid \mathbf{0} \right)_{l \times (H-l)} \).
\]
As a result, \( \mathbf{V}_s = \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \otimes \left( \mathbf{I}_l - \frac{\mathbf{J}_l}{H} \right) \).

Also, \( \mathbf{V}_{sr} = \mathbf{K}_{11} \text{var}_s (\mathbf{Y}) \mathbf{K}_{12} \) and
\[
\mathbf{K}_{12} = \mathbf{I}_H \otimes \left( \mathbf{0} \mid \mathbf{I}_{H-l} \right) \text{. As a result, } \mathbf{V}_{sr} = \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \otimes \left( \frac{\mathbf{J}}{H} \right) \text{.}
\]

We wish to obtain explicit expressions for the estimator. From Corollary 18.2.10 (Harville, 1997 #1031)(p424), \( \mathbf{I}_a - \frac{1}{k} \mathbf{J}_a \)\(^{-1} = \mathbf{I}_a + \frac{1}{k-a} \mathbf{J}_a \) when \( a \neq k \). Thus,
\[
\mathbf{V}_s^{-} = \left[ \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \right]^{-} \left( \mathbf{I}_l + \frac{1}{H-1} \mathbf{J}_l \right) \text{.}
\]
Then, since \( \mathbf{X}_s = \mathbf{I}_H \otimes \frac{\mathbf{l}}{H} \),
\[
\left( \mathbf{X}_s \mathbf{V}_s^{-} \mathbf{X}_s \right) = \left[ \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \right]^{-} \left( \mathbf{I}_l + \frac{1}{H-1} \mathbf{J}_l \right) \frac{\mathbf{l}}{H} \text{, which simplifies to}
\]
\[
\left( \mathbf{X}_s \mathbf{V}_s^{-} \mathbf{X}_s \right) = \frac{I(H-1)}{H(H-1)} \left[ \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \right]^{-} \left( \mathbf{I}_l + \frac{1}{H-1} \mathbf{J}_l \right) \frac{\mathbf{l}}{H} \text{.}
\]
We use this expression to evaluate \( \left( \mathbf{X}_s \mathbf{V}_s^{-} \mathbf{X}_s \right) = \frac{H(H-1)}{I(H-1)} \left[ \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \right]^{-} \).

Also, since \( \mathbf{X}_s \mathbf{V}_s^{-} = \left[ \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \right]^{-} \left( \mathbf{I}_l + \frac{1}{H-1} \mathbf{J}_l \right) \text{ simplifies to}
\]
\[
\mathbf{X}_s \mathbf{V}_s^{-} = \left[ \frac{1}{(H-1)} \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{H} \right) \mathbf{D}_y \right]^{-} \frac{\mathbf{l}}{H-1} \text{, and}
\]
\[ V_s^{-1}X_s = \left[ \frac{1}{(H-1)} D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \frac{I_l}{H-I} \]

\[ V_s^{-1}X_s \left( X_s'V_s^{-1}X_s \right)^{-1} = \frac{H}{I} \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes I_l \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \]

Also,

\[ V_s^{-1}X_s \left( X_s'V_s^{-1}X_s \right)^{-1} = \frac{H}{I} \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes I_l \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes I_l \]

From properties of generalized inverses (Harville, 1997 #1031) (p107),

\[ \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes I_l \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] = \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \]

hence

\[ V_s^{-1}X_s \left( X_s'V_s^{-1}X_s \right)^{-1} = \frac{H(H-1)}{I(H-I)} \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes J_l \]

We use this expression to evaluate \( V_s^{-1} = V_s^{-1}X_s \left( X_s'V_s^{-1}X_s \right)^{-1}X_s'V_s^{-1} \). Since

\[ V_s^{-1} = \left[ \frac{1}{(H-1)} D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \left( I_l + \frac{1}{H-I} J_l \right) \]
\[ V_s - V_s X_s \left( X_s' V_s X_s \right)' X_s' V_s = \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \left( (H-1) I + \frac{H-1}{H-I} J \right) \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \frac{H(H-1)}{I(H-I) J} \]

which simplifies to

\[ V_s - V_s X_s \left( X_s' V_s X_s \right)' X_s' V_s = \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes (H-1) \left( I - \frac{1}{I} J \right). \]

We complete the evaluation of an expression for \( L' \) by evaluating \( V_s L, X_s \) and \( X'_s L_s \).

Also, we complete the evaluation of the expression for \( L'_s \) by evaluating \( V_s L, 1_H \) and

\[ X'_s L_s 1_H. \] Now \( V_s X = \frac{1}{(H-1)} D_y \left( I_H - \frac{J_H}{H} \right) D_y \otimes \left( -\frac{J_H}{H} \right) \) and \( L_s = I_H \otimes 1_{H-I}. \) As a result, \( V_s L_s = \frac{1}{(H-1)} D_y \left( I_H - \frac{J_H}{H} \right) D_y \otimes \left( -\frac{J_H}{H} \right) \) which simplifies to

\[ V_s L_s = D_y \left( I_H - \frac{J_H}{H} \right) D_y \otimes \left( -\frac{(H-I)}{H} \right) \]

while

\[ V_s L_s 1_H = D_y \left( I_H - \frac{J_H}{H} \right) y \otimes \left( -\frac{(H-I)}{H} \right) \]

Also, since \( X_s = I_H \otimes \frac{1_{H-I}}{H} \),

\[ X'_s L_s = \left( I_H \otimes \frac{1_{H-I}}{H} \right) (I_H \otimes \frac{1_{H-I}}{H}) = \frac{H-I}{H} \]

while \( X'_s L_s 1_H = \frac{H-I}{H} \).
Finally, combining terms, since an estimator of \( L_{rs} \) is given by

\[
L_{rs} = \left( V_s - V_s X_s \left( X_s^\prime V_s X_s \right) \right) X_s^\prime V_s \left( X_s^\prime V_s X_s \right)^\prime \left( V_s^\prime L_y + V_s X_s \left( X_s^\prime V_s X_s \right)^\prime \right) X_s^\prime, \] 

then

\[
L_{rs} = \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes (H-1) \left( I_L - \frac{1}{I} J_I \right) \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \left( \frac{(H-I)}{(H-1)} \frac{1}{I} \right) \] 

\[+ \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \frac{H-I}{I} I_l \] 

which simplifies to

\[
L_{rs} = \left( \frac{H-I}{I} \right) \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes I_l \] 

in this expression is zero, and hence

\[
L_{rs} = \left( \frac{H-I}{I} \right) \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes I_l \] 

In a similar manner the estimate of

\[
L_{rs*} = \left( V_s - V_s X_s \left( X_s^\prime V_s X_s \right) \right) \left( V_s^\prime L_y + V_s X_s \left( X_s^\prime V_s X_s \right) \right) \left( X_s^\prime, L_y, H \right) \] 

is given by

\[
L_{rs*} = \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes (H-1) \left( I_L - \frac{1}{I} J_I \right) \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \left( \frac{(H-I)}{(H-1)} \frac{1}{I} \right) \] 

\[+ \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \left[ D_y \left( I_H - \frac{J_H}{H} \right) D_y \right] \otimes \frac{H-I}{I} I_l \] 

which simplifies to
\[
L_{r^*} = \left[ \begin{bmatrix} \sum \{ D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \} \right] D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) y \right] \otimes \left( \frac{(H-I)}{H} \right) \left( \mathbf{I}_H - \frac{1}{I} J_I \right) \mathbf{i}_I \\
+ \left( \frac{H-I}{I} \right) \left[ D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \right] \left( D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \right) \otimes \mathbf{I}_H \otimes \mathbf{i}_I \right].
\]

The first term in this expression is zero, and hence

\[
L_{r^*} = \left( \frac{H-I}{I} \right) \left[ D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \right] \left( D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \right) \otimes \mathbf{I}_H \otimes \mathbf{i}_I).
\]

### 3.7.2.1. Aside on Simplifications of Generalized Inverses

Suppose that none of the elements of \( y \) are zero. In this case,

\[
\left[ D_y \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \right] = D_y - \left( \mathbf{I}_H - \frac{J_H}{H} \right) D_y \text{, where } D_y^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \frac{1}{y_1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.
\]

We evaluate the generalized inverse of \( \mathbf{A} = \left( \mathbf{I}_H - \frac{J_H}{H} \right) \) making use of Theorem 9.2.7 (p116)(Harville, 1997 #1031). This theorem says that if \( \mathbf{A} \) is an \( H \times H \) matrix, and \( \mathbf{G} \) is any particular generalized inverse of \( \mathbf{A} \), then \( \mathbf{G}^* \) is a generalized inverse of \( \mathbf{A} \) if and only if

\[
\mathbf{G}^* = \mathbf{G} + (\mathbf{I}_H - \mathbf{GA}) \mathbf{T} + \mathbf{S}(\mathbf{I}_H - \mathbf{AG}) \text{ for some } H \times H \text{ matrices } \mathbf{S} \text{ and } \mathbf{T}.
\]

Now let \( \mathbf{G} = \begin{pmatrix} \mathbf{A}_{11}^{-1} \cdot 0 \\ \begin{bmatrix} \mathbf{A}_{11}^{-1} \end{bmatrix} \end{bmatrix} \right) \) where \( \mathbf{A}_{11} = \begin{pmatrix} \mathbf{I}_{H-1} - \frac{J_{H-1}}{H} \end{bmatrix} \), and \( \mathbf{A}_{11}^{-1} = \begin{pmatrix} \mathbf{I}_{H-1} + \frac{J_{H-1}}{H} \end{bmatrix} \).

Now \( \mathbf{GA} = \left( \mathbf{I}_{H-1} + \frac{J_{H-1}}{H} \right) \begin{pmatrix} \mathbf{A}_{11}^{-1} \cdot 0 \\ \begin{bmatrix} \mathbf{A}_{11}^{-1} \end{bmatrix} \end{bmatrix} \right) \) which can be expressed as

\[
\begin{pmatrix} \mathbf{I}_H - \frac{J_H}{H} \end{pmatrix}
\]
\[ \mathbf{GA} = \left( \begin{array}{cc} \mathbf{I}_{H-1} + \mathbf{J}_{H-1} & 0 \\ 0 & 0 \end{array} \right) \frac{\mathbf{H} - \mathbf{0}}{\mathbf{I}_b \times \mathbf{1}_b} - \frac{1}{\mathbf{H}} \left( \begin{array}{cc} \mathbf{J}_{H-1} & \mathbf{1}_{H-1} \\ \mathbf{0} & \mathbf{0} \end{array} \right) - \frac{\mathbf{H} - \mathbf{1}}{\mathbf{H}} \left( \begin{array}{cc} \mathbf{J}_{H-1} & \mathbf{1}_{H-1} \\ \mathbf{0} & \mathbf{0} \end{array} \right), \text{ and simplifies to} \]

\[ \mathbf{GA} = \left( \begin{array}{cc} \mathbf{I}_{H-1} + \mathbf{J}_{H-1} & 0 \\ 0 & 0 \end{array} \right) \frac{\mathbf{H} - \mathbf{0}}{\mathbf{I}_b \times \mathbf{1}_b} - \frac{1}{\mathbf{H}} \left( \begin{array}{cc} \mathbf{J}_{H-1} & \mathbf{1}_{H-1} \\ \mathbf{0} & \mathbf{0} \end{array} \right) = \left( \begin{array}{cc} \mathbf{I}_{H-1} & -\mathbf{1}_{H-1} \\ \mathbf{0} & \mathbf{0} \end{array} \right). \text{ Then} \]

\[ (\mathbf{I}_H - \mathbf{GA}) = \left( \begin{array}{cc} \mathbf{0} & \mathbf{1}_{H-1} \\ \mathbf{H} \times \mathbf{1}_b & -\mathbf{1}_{H-1} \end{array} \right). \text{ In a similar manner,} \]

\[ \mathbf{AG} = \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) \left( \begin{array}{cc} \mathbf{I}_{H-1} + \mathbf{J}_{H-1} & 0 \\ 0 & 0 \end{array} \right) \text{ simplifies to} \]

\[ \mathbf{AG} = \left( \begin{array}{cc} \mathbf{I}_{H-1} & \mathbf{0} \\ \mathbf{H} \times \mathbf{1}_b & \mathbf{0} \end{array} \right), \text{ so that} \]

\[ (\mathbf{I}_H - \mathbf{AG}) = \left( \begin{array}{cc} \mathbf{0} & \mathbf{1}_{H-1} \\ \mathbf{H} \times \mathbf{1}_b & -\mathbf{1}_{H-1} \end{array} \right). \text{ Combining terms,} \]

\[ \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) = \left( \begin{array}{cc} \mathbf{I}_{H-1} + \mathbf{J}_{H-1} & 0 \\ 0 & 0 \end{array} \right) + \left( \begin{array}{cc} \mathbf{0} & \mathbf{1}_{H-1} \\ \mathbf{H} \times \mathbf{1}_b & -\mathbf{1}_{H-1} \end{array} \right) \text{ in a simple example.} \]

3.7.2.2. A Simple Example

Our goal is to evaluate an explicit expression for

\[ \mathbf{L}_{r^*} = \left( \begin{array}{c} \frac{\mathbf{H} - \mathbf{I}}{\mathbf{I}} \end{array} \right) \left[ \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) \mathbf{D}_y \right] \left[ \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) \mathbf{D}_y \right] \right) \times \mathbf{1}_i \text{ in a simple example.} \]

Suppose that none of the elements of \( y \) are zero. Then

\[ \mathbf{L}_{r^*} = \frac{\mathbf{H} - \mathbf{I}}{\mathbf{I}} \left[ \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) \mathbf{D}_y \right] \left[ \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) \mathbf{D}_y \right] \times \mathbf{1}_i \]

\[ = \left( \frac{\mathbf{H} - \mathbf{I}}{\mathbf{I}} \right) \left[ \mathbf{D}_y \left( \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right) \left[ \mathbf{I}_H - \frac{\mathbf{J}_H}{\mathbf{H}} \right] \mathbf{D}_y \right] \times \mathbf{1}_i \]
Also, let us take \( \left( I_{H} - \frac{J_{H}}{H} \right)^{-1} = \begin{pmatrix} (I_{H-1} - \frac{1}{H} J_{H-1})^{-1} & 0 \\ 0 & 0 \end{pmatrix} \), and

\[
\begin{pmatrix}
(I_{H-1} + J_{H-1}) & 0 \\
0 & 0
\end{pmatrix}^{-1} = \begin{pmatrix}
(I_{H-1} + J_{H-1})^{-1} & 0 \\
0 & 0
\end{pmatrix} = \begin{pmatrix}
(I_{H-1} - \frac{1}{H} J_{H-1}) & 0 \\
0 & 0
\end{pmatrix}.
\]

Then

\[
L_{rs}^{o} = \left( \frac{H - I}{I} \right) \left( D_{y}^{-1} \begin{pmatrix}
(I_{H-1} + J_{H-1}) & 0 \\
0 & 0
\end{pmatrix} \begin{pmatrix}
(I_{H-1} - \frac{1}{H} J_{H-1}) & 0 \\
0 & 0
\end{pmatrix} D_{y} \right) \otimes 1_{L}.
\]

Then

\[
\begin{pmatrix}
(I_{H-1} + J_{H-1}) & 0 \\
0 & 0
\end{pmatrix}^{-1} = \begin{pmatrix}
(I_{H-1} - \frac{1}{H} J_{H-1}) & 0 \\
0 & 0
\end{pmatrix}.
\]

Then

\[
L_{rs}^{o} = \left( \frac{H - I}{I} \right) \left( D_{y}^{-1} \begin{pmatrix}
(I_{H-1}) & 0 \\
0 & 0
\end{pmatrix} D_{y} \right) \otimes 1_{L}
\]

which simplifies to

\[
L_{rs}^{o} = \left( \frac{H - I}{I} \right) \otimes 1_{L}, \text{ or } L_{rs}^{o} = \left( \frac{H - I}{I} \right) \otimes 1_{L}.
\]

\[3.7.2.3. \text{ A Simple Application}\]

We use this expression to evaluate estimates of \( \beta \). Recall that\( \hat{\beta} = L_{s}^{\prime} Y_{s} + L_{rs}^{\prime} Y_{s} \),

where \( L_{s}^{\prime} = I_{H} \otimes 1_{L} \) and \( L_{rs}^{\prime} = \left( \frac{H - I}{I} \right) \left( I_{H-1} \otimes 1_{L} \right) \). Adding these two terms together, we find that
\[
L_s' + L_{rs}' = I_H \otimes 1_i' + \left( \frac{H-I}{I} \right) \begin{pmatrix}
  I_{H-1} & 0 \\
  0 & 1
\end{pmatrix} \otimes 1_i'.
\]

Recall that \( Y_s' = \left( \begin{array}{c}
(Y_{11} \cdots Y_{1i}) \\
(Y_{21} \cdots Y_{2i}) \\
\vdots
\end{array} \right) \) and that the estimator of \( \beta = (L_s' + L_{rs}')Y_s \) is given by

\[
\hat{\beta} = \left( \frac{H}{I} \sum_{i=1}^{l} Y_{1i} \right)
\]

\[
\left( \frac{H}{I} \sum_{i=1}^{l} Y_{2i} \right)
\]

\[
\vdots
\]

\[
\left( \sum_{i=1}^{l} Y_{si} \right)
\]

For instance, suppose that \( H = 5 \), \( I = 3 \), and \( y' = (2 \quad 7 \quad 33 \quad 18 \quad 4) \), and that the selected units in the sample correspond to the units \( (i = 1, s = 2); (i = 2, s = 5); (i = 3, s = 1) \).

Then \( \hat{\beta} = \left( \frac{5}{3} \sum_{i=1}^{3} Y_{1i} \right) \) or equivalently, \( \hat{\beta} = \left( \frac{5}{3} \right) \begin{pmatrix}
2 \\
7 \\
33 \\
18 \\
4
\end{pmatrix} = \left( \begin{array}{c}
3.33 \\
11.67 \\
0 \\
0 \\
4
\end{array} \right) \).
3.7.2.4. Discussion of the Example

At first sight, this estimator doesn’t appear to make sense. Rather than using the value of the unit observed, we calculate an inflated value. For example, since in the sample the response for unit $s=1$ was observed to be equal to 2, we know the value for the unit to be equal to 2. Nevertheless, the estimator of the value for the unit is given by 3.33. This may not appear to make any sense. This seems to make no sense since when a PSU is realized, the observed value for the PSU is the parameter we are after. However, it is possible that the estimator we obtain is consistent with the estimation framework. We constructed the estimator not conditional on selection of a set of PSUs, but unconditionally. Hence the properties of the estimator need to be evaluated over all samples, including samples where the PSU may not be realized in the sample. In our example, the first PSU (ie. $s=1$) will be in $3/5$ of all possible samples. For these samples, the estimate of the PSU $s=1$ parameter is $\hat{\beta}_1 = \frac{5}{3} y_1$. For the remaining $2/5$ of the samples, the estimate of the PSU $s=1$ parameter is $\hat{\beta}_1 = 0$. Averaging the values of these estimators over all samples, we find

$$\frac{3\left(\frac{5}{3} y_1\right) + 2(0)}{5} = y_1.$$ 

Thus, our estimator is an unbiased estimate of the parameter for $s=1$. This is what we set out to obtain.

There is an exception to this unbiased property of the estimator that occurs for the last unit ($s = H$). This is related to the particular choice of the generalized inverse in the estimator. It may be possible to choose a generalized inverse that makes expressions for unit estimators identical.
A final comment concerns the weight that is attached to the unit’s estimator. This weight is equal to one over the inclusion probability for the unit, or an inverse sampling probability weight. It is possible (although we have not investigated it here) that if we considered a superpopulation model with unequal probability sampling, that the estimator weights would once again be one over the inclusion probability. This may relate to the Horvitz-Thompson estimator. Note that as the sample size approaches the population size, the weight approaches one. If the weight is one, then the estimator is simply the observed value for the unit. Traditionally, such an estimator would be called a fixed effect.

4. Summary of Estimation Strategy

We summarize the steps in formulating the model and constructing the estimator as follows:

1. Specify the model: \( Y = X \beta + E \).

2. Choose \( X \) to equate population and superpopulation parameters: \( L'X = I \) where \( L'Y = \beta \).

3. Re-arrange the superpopulation into the sample and remainder via the permutation matrix \( K \) resulting in \( \begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix} \beta + \begin{pmatrix} E_s \\ E_r \end{pmatrix} \).

4. Require the estimator given by \( L_s \) to be linear in data, and unbiased implying that \( L_s'X_s' = L_r'X_r' \).

5. Estimate \( \beta \) by \( \hat{\beta} = L_s'Y_s + L_r'Y_s \) such that the GMSE is minimized subject to the constraint in (4) where
GMSE = \mathbf{1}_H' \mathbb{E} \left[ (\hat{\beta} - \beta)(\hat{\beta} - \beta)' \right] \mathbf{1}_H = \mathbf{1}_H' \left( \mathbf{L}_{rs}' \begin{bmatrix} -\mathbf{L}_r' \end{bmatrix} \right) \text{var} \left( \begin{bmatrix} \mathbf{Y}_z \mathbf{L}_{rs} \\ -\mathbf{Y}_r \end{bmatrix} \right) \mathbf{1}_H.'
References


Neyman, (1923)


