Research Article **Open Access**

Stephen Haslett*

Best linear unbiased estimation for varying probability with and without replacement sampling

https://doi.org/110.1515/spma-2019-0007
Received April 14, 2018; accepted August 22, 2018

Abstract: When sample survey data with complex design (stratification, clustering, unequal selection or inclusion probabilities, and weighting) are used for linear models, estimation of model parameters and their covariance matrices becomes complicated. Standard fitting techniques for sample surveys either model conditional on survey design variables, or use only design weights based on inclusion probabilities essentially assuming zero error covariance between all pairs of population elements. Design properties that link two units are not used. However, if population error structure is correlated, an unbiased estimate of the linear model error covariance matrix for the sample is needed for efficient parameter estimation. By making simultaneous use of sampling structure and design-unbiased estimates of the population error covariance matrix, the paper develops best linear unbiased estimation (BLUE) type extensions to standard design-based and joint design and model based estimation methods for linear models. The analysis covers both with and without replacement sample designs. It recognises that estimation for with replacement designs requires generalized inverses when any unit is selected more than once. This and the use of Hadamard products to link sampling and population error covariance matrix properties are central topics of the paper. Model-based linear model parameter estimation is also discussed.

Keywords: Best linear unbiased estimator; best linear unbiased predictor; generalized inverse; Hadamard product; linear models; positive semidefiniteness; sample surveys; survey design; sampling with replacement; sampling without replacement; superpopulation.

1 Introduction

There are two relatively distinct methodologies for analysis of sample survey data collected via a complex sampling scheme that may include stratification, clustering and weighting of responses.

In the 1980s and early 1990s there was considerable academic debate around whether design-based or model-based methods were better. In time there was a rapprochement. Model-assisted sampling was developed in Särndal et al. [17] . This had the major benefit of making explicit the underlying models used by practitioners to determine inclusion (or selection) and joint inclusion probabilities when determining a good design-based survey design. The model-assisted analysis of sample surveys had been foreshadowed, for example in Cochran [3] and the joint design- and model-based approach was considered in detail in Haslett [8] and used in Fuller [4].

Design-based methods are based on inclusion probability (the probability each unit is included in the sample) and joint inclusion probabilities (the set of joint probabilities that two given units are included in the sample). An alternative for with replacement sampling which is considered in detail in this paper, is based

***Corresponding Author: Stephen Haslett:** Massey University, New Zealand and The Australian National University, Australia. E-mail: s.j.haslett@massey.ac.nz, stephen.haslett@anu.edu.au

on selection probabilities (the probability a unit is selected at each draw). Usually, design-based frameworks are used to estimate descriptive statistics such as means, totals and their variances. For linear models fitted to sample survey data, design-based regression parameter estimation allows for stratification and clustering and makes use of inverse inclusion probabilities for individual units as weights, but usually ignores joint inclusion probabilities for pairs of units. One reason for this simplification is that otherwise, if the error covariance in the linear model is not a diagonal matrix and joint inclusion probabilities are incorporated to improve estimation, there are complications that cannot be resolved using matrix multiplication alone. The sample-based unbiased estimator of the $N \times N$ covariance matrix is then the Hadamard product of the elementwise inverse of inclusion probabilities and joint inclusion probabilities, the population covariance matrix, and an inclusion matrix for the sample of rank equal to the sample size, n. However, even the $n \times n$ submatrix of the non-zero rows and columns of this Hadamard product which correspond to the sampled elements is not always positive definite. For linear models that include survey design information, this has important consequences for best linear unbiased estimation (BLUE) for both with and without replacement sampling.

Conditions under which a change in covariance structure leaves BLUEs unchanged are given in Rao [14]. An extension to best linear unbiased prediction (BLUP), and/or to BLUE is outlined in Haslett & Puntanen [9]. Surprisingly perhaps, the class of "equivalent" matrices for linear models contains matrices that are symmetric but not positive semidefinite. These results can be used to explore how, while retaining the same BLUEs and/or BLUPs, covariances estimated from the sample used in design-based estimation might be adjusted to meet the requirement for positive semidefiniteness.

An alternative view on analysis of sample surveys is model-based. An early reference is Royal & Cumberland [16]. See also Chambers & Clark [1]. For linear models and complex samples, model-based analysis usually includes design information via supplementary auxiliary variables, so that inference on the other associated parameters in the model are conditional on the design. Clustering is accounted for via non-diagonal covariance matrices between population and sampled units. For example, clustering is often incorporated by having equal covariances between units in the same cluster, but zero correlation between units in different clusters. Neither selection nor inclusion and joint inclusion probabilities are generally used, and parameter estimates via design- and model-based methods are not necessarily equal.

2 Design-based and model-based estimation for linear models

Returning to the design-based framework, suppose for a given population of size N that the population mean of the variable

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Let $X_i = 1$ if $i \in s$ where s denotes the sample, and $X_i = 0$ if $i \notin s$. Then $X_i = 1$ for n of the population units, and is zero for the remaining $N-n$ unsampled units. In the design-based context, $Y_i = y_i$ for $\{i = 1, 2, \ldots, N\}$ and the only random variables are the $\{X_i : i = 1, 2, \ldots, N\}$. The only design-unbiased estimator of the mean is then:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} X_i y_i/\pi_i$$

with $\mathrm{E}(X_i) = \pi_i$ for $i = 1, 2, \ldots, N$. Here E is the design expectation and π_i is the inclusion probability for the ith unit. This is called the Horvitz-Thompson estimator (Horvitz & Thompson [10]).

In model-based methods, the population is taken to be a sample from a superpopulation with model-expectation E and $N \times N$ model-covariance matrix $\boldsymbol{V}$. Each member of $\{Y_i : i = 1, 2, \ldots, N\}$ is random with respect to the superpopulation, and so is $\bar{Y}$. When both design and model-based methods are integrated, then design-based, model-based, model-assisted estimators, and joint design and model-based estimators, plus their variances and estimated variances can be derived. Further details can be found in Fuller [4] or Haslett [8]. This broader framework is important because then time series, as well as linear and generalized linear

models can be fitted to sample survey data to provide better estimates of the parameters and of their covariance matrices. The standard, design-based way to fit a linear model to survey data is to use inverse inclusion probabilities as weights within a model-based context. The linear model for the survey data is specified as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} \tag{2.1}$$

Here $\boldsymbol{Y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ matrix of auxiliary variables, $\boldsymbol{\beta}$ a $p \times 1$ vector of parameters, and $\boldsymbol{e}$ an $n \times 1$ vector of errors with covariance matrix $\boldsymbol{V}(\boldsymbol{e}) = E(\boldsymbol{ee}') = \boldsymbol{V_e}$. Neither $\boldsymbol{X}$ nor $\boldsymbol{V_e}$ need be of full rank. Generally $\boldsymbol{V_e}$ is unknown and is estimated by $\widehat{\boldsymbol{V_e}}$.

For the sample survey data, $\boldsymbol{Y} = (y_1, y_2, \ldots, y_i, \ldots, y_n)'$ the inclusion probabilities are $\boldsymbol{\Pi_0} = \text{diag}(\pi_i)$ where $i = 1, 2, \ldots, n$. The notation is also extended to include the $N - n$ non-sampled elements. All N inclusion probabilities $\{\pi_i : i = 1, 2, \ldots, N\}$ and $N(N+1)/2$ possibly different joint inclusion probabilities $\{\pi_{ij} : i = 1, 2, \ldots, N; j = 1, 2, \ldots, N\}$ where $\pi_{ij} = \pi_{ji}$ and $\pi_{ii} = \pi_i$ by definition, are specified at design stage.

The standard design-based least squares solution for full rank $\boldsymbol{X}$ (e.g., Chambers & Skinner [2], Skinner et al. [18]) is then

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Pi_0^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Pi_0^{-1}}\boldsymbol{Y} \tag{2.2}$$

with estimated covariance matrix given by

$$\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\beta}}) = \{(\boldsymbol{X}'\boldsymbol{\Pi_0^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Pi_0^{-1}}\}\widehat{\boldsymbol{V_e}}\{\boldsymbol{\Pi_0^{-1}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Pi_0^{-1}}\boldsymbol{X})^{-1}\} \tag{2.3}$$

This design-based solution (2.2) is weighted least squares, using weights equal to the inverse of the inclusion probabilities for the sampled units. However it also has a connection to ordinary least squares in that the model covariance $\boldsymbol{V_e}$ is not involved. However there is an adjustment for an estimate of $\boldsymbol{V_e}$ in the estimated covariance matrix $\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\beta}})$, so that (2.3) is not necessarily equal to the appropriately scaled version of $(\boldsymbol{X}'\boldsymbol{\Pi_0^{-1}}\boldsymbol{X})^{-1}$ that is the estimated covariance matrix of $\tilde{\beta}$ for simple random sampling.

3 Linear models adjusted for inclusion and joint inclusion probabilities and for covariance structure

A population of size N can be considered as being sampled from a superpopulation, and the linear model based on that population is

$$\boldsymbol{Y}_P = \boldsymbol{X}_P\boldsymbol{\beta} + \boldsymbol{e}_P\,, \tag{3.1}$$

with $E(\boldsymbol{e}_P) = \boldsymbol{0}$, where E denotes superpopulation expectation, and $\boldsymbol{V}(\boldsymbol{e}_P) = E(\boldsymbol{e}_P\boldsymbol{e}_P{}') = \boldsymbol{V_{e_P}}$. is $N \times N$ with (i, j)th element v_{ij}. Then the best linear unbiased estimate (BLUE) of the superpopulation parameter $\boldsymbol{\beta}$ in the full rank case is

$$\widehat{\boldsymbol{\beta}_P} = (\boldsymbol{X}_P{}'\boldsymbol{V_{e_P}^{-1}}\boldsymbol{X}_P)^{-1}\boldsymbol{X}_P{}'\boldsymbol{V_{e_P}^{-1}}\boldsymbol{Y}_P \tag{3.2}$$

3.1 Design unbiased estimation of the population covariance matrix

Suppose we have a sample s of size n which has been selected from the population P with inclusion probabilities $\{\pi_i : i = 1, 2, \ldots, N\}$ and joint inclusion probabilities $\{\pi_{ij} : i = 1, 2, \ldots, N; j = 1, 2, \ldots, N\}$.

Define $\boldsymbol{\chi}_P$ to be the $N \times N$ matrix with ijth element $\chi_{ij} = 1$ if both $i \in s$ and $j \in s$ and zero otherwise. $\boldsymbol{\chi}_P$ varies depending on the sample drawn, has n non-zero diagonal elements all equal to one, and $n(n-1)$ off-diagonal elements each equal to one; all other $N^2 - n^2$ elements are zero.

Provided the design is noninformative, $\text{E}(\boldsymbol{\chi}_\text{P}) = \boldsymbol{\Pi_P}$ where E is expectation with respect to the design, and $\boldsymbol{\Pi_P}$ has ijth element π_{ij} for $i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$. The key property of a noninformative sample design is that selection and joint inclusion probabilities are independent of the errors $\boldsymbol{e}_P$ in (3.1). Note that all elements of $\boldsymbol{\Pi_P}$ are positive so $\boldsymbol{\Pi_P}$ is a positive matrix, and since no element of $\boldsymbol{\chi}_P$ is less than zero $\boldsymbol{\chi}_P$ is a non-negative matrix.

Denote the Hadamard inverse of $\boldsymbol{\Pi_P}$ by $\boldsymbol{\Pi_P}^{\odot -}$ so that $\boldsymbol{\Pi_P}^{\odot -} \odot \boldsymbol{\Pi_P} = \mathbf{11'}$ where $\mathbf{1}$ is an $N \times 1$ vector of ones, i.e., each element of $\boldsymbol{\Pi_P}^{\odot -}$ equals the inverse of the corresponding element of $\boldsymbol{\Pi_P}$. Then from Haslett [7] we have:

Theorem 3.1. *The expected value of the augmented weighted sample superpopulation variance,*

$$\boldsymbol{V}_{\mathbf{e}_P,s} = \boldsymbol{\chi}_P \odot \boldsymbol{V}_{\mathbf{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -} \tag{3.3}$$

is

$$\mathrm{E}(\boldsymbol{V}_{\mathbf{e}_P,s}) = \mathrm{E}(\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\mathbf{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}) = \boldsymbol{V}_{\mathbf{e}_P} \tag{3.4}$$

where $\odot$ denotes the Hadamard (or elementwise) product, $\boldsymbol{V}_{\mathbf{e}_P}$ is positive semidefinite, $\boldsymbol{\chi}_P$ is a non-negative matrix, and $\boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ is a positive matrix.

Proof: $\pi_{ij} > 0$ for all $i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$, so that all entries of $\boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ are positive. So we can define the Hadamard or elementwise inverse of $\boldsymbol{\Pi_P}$, namely the $N \times N$ matrix $\boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ to have ijth element $1/\pi_{ij}$. Similarly $\boldsymbol{\chi}_P$ is a non-negative matrix because all its elements are either one or zero. Then for $i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$, we have $\mathrm{E}(\chi_{ij} v_{ij} \pi_{ij}^{-1}) = v_{ij} \pi_{ij}^{-1} \mathrm{E}(\chi_{ij}) = v_{ij} \pi_{ij}^{-1} \pi_{ij} = v_{ij}$ which in matrix form is (3.3).

Note that $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ contains only n of N rows and n of N columns that are non-zero, and that $\boldsymbol{V}_{\boldsymbol{e}_P}$ is assumed known. So, after suitable permutation, there is only an $n \times n$ submatrix of $\boldsymbol{\chi}_P$ that is non-zero; $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ has ijth element v_{ij}/π_{ij} if $i \in s$ and $j \in s$ and is zero otherwise, with the convention that the diagonal elements are v_{ii}/π_i if $i \in s$ and zero otherwise, and hence (after the same permutation) $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ also contains only an $n \times n$ submatrix that is non-zero.

Note too that if $\boldsymbol{V}_{\boldsymbol{e}_P} = \sigma_{e_P}^2 \boldsymbol{I}$, where $\sigma_{e_P}^2$ is a scale factor, then $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ reduces to a diagonal matrix with ith diagonal element v_{ii}/π_i if $i \in s$ and 0 if $i \notin s$. If all members of $\{v_{ii} : i = 1, 2, \ldots, N\}$ are equal, then $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -} = \sigma_{e_P}^2 \left\{\mathrm{diag}(\pi_i)_{N \times N}\right\}^{-1}$ which for the sample corresponds to the standard case in (2.2).

3.2 Improved approximation to BLUE for design-based sample survey estimators

Permuting both rows and corresponding columns of $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ so that the sampled elements occur in the first n rows and n columns is straightforward. Denote this re-ordering by $\boldsymbol{V}_{\boldsymbol{e}_P,s}$ and its submatrix made up of the first n rows and n columns by $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$. Now under a full rank condition on $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$, the $N \times N$ matrix $\boldsymbol{V}_{\boldsymbol{e}_P,s}$ has a generalized inverse (which is also the Moore-Penrose inverse)

$$\boldsymbol{V}_{\boldsymbol{e}_P,s}^{+} = \begin{pmatrix} \boldsymbol{V}_{\boldsymbol{e}_P,s,n}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{3.5}$$

where $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ has ijth element v_{ij}/π_{ij}. Because $0 < \pi_i \leq 1$ for $i = 1, 2, \ldots, N$, the inverse $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}^{-1}$ exists for all possible samples provided $\boldsymbol{V}_{\boldsymbol{e}_P}$ is full rank. Note that for simplicity of notation, once sampled there is an implicit relabelling of the units so that those in the selected sample are relabelled (in the same order) as $i = 1, 2 \ldots, n$ and $j = 1, 2, \ldots, n$. After the rows of $\boldsymbol{X}_P$ have also been appropriately permuted to match the permutation for the rows and columns of $\boldsymbol{V}_{\boldsymbol{e}_P,s}^{+}$:

$$\boldsymbol{X}_P{}' \boldsymbol{V}_{\boldsymbol{e}_P,s}^{+} \boldsymbol{X}_P = \begin{pmatrix} \boldsymbol{X}' & \boldsymbol{X}_{\sim s}{}' \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_{\boldsymbol{e}_P,s,n}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}_{\sim s} \end{pmatrix} = \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{e}_P,s,n}^{-1} \boldsymbol{X} \tag{3.6}$$

and

$$\boldsymbol{X}_P{}' \boldsymbol{V}_{\boldsymbol{e}_P,s}^{+} \boldsymbol{Y}_P = \begin{pmatrix} \boldsymbol{X}' & \boldsymbol{X}_{\sim s}{}' \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_{\boldsymbol{e}_P,s,n}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Y}_{\sim s} \end{pmatrix} = \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{e}_P,s,n}^{-1} \boldsymbol{Y} \tag{3.7}$$

where, for the non-sampled elements, $\boldsymbol{Y}_{\sim s}$ denotes the y-values and $\boldsymbol{X}_{\sim s}$ contains the auxiliary variables.

Thus from (3.6) and (3.7), the approximate BLUE based on the sampled elements is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{Y} \tag{3.8}$$

with estimated covariance matrix given by

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = \{(\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\}\widehat{\boldsymbol{V}_{\boldsymbol{e}}}\{\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{X})^{-1}\} \tag{3.9}$$

where, as in Section 2, $\boldsymbol{V}(\boldsymbol{e}) = \boldsymbol{V}_{\boldsymbol{e}}$. and $\boldsymbol{e}$ denotes that part of $\boldsymbol{e}_P$ that corresponds to the n sampled elements. When $\boldsymbol{V}_{\boldsymbol{e}_P}$ is diagonal, i.e., $\boldsymbol{V}_{\boldsymbol{e}_P} = \sigma^2_{e_P}\boldsymbol{I}$, then $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ is also diagonal and (3.8) and (3.9) reduce to (2.2) and (2.3) respectively.

One major advantage of (3.8) and (3.9) over (2.2) and (2.3) is that they can be applied to estimation of fixed effects in mixed linear models, where incorporation of the random effects into $\boldsymbol{V}_{\boldsymbol{e}_P}$ means that it is no longer a diagonal matrix, so that (2.2) and (2.3) cannot be applied.

3.3 Is the augmented weighted sample superpopulation variance positive semidefinite, and is this necessary for BLUE?

Perhaps surprisingly, following from Haslett & Puntanen [9] and the earlier results in Rao [14], to produce the correct estimates of β from (3.6), $\boldsymbol{V}_{\boldsymbol{e}_P,s}$, the $(N \times N)$ augmented weighted sample superpopulation variance need not be positive semidefinite, and its $n \times n$ submatrix, $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ need not be positive definite. See also Haslett [7]. This can be seen for fixed effect linear models from an extension to Rao [14]. Given a linear model of the form (2.1) with error covariance matrix $\boldsymbol{V}_1$, then for any $\boldsymbol{V}_2$ of the form

$$\boldsymbol{V}_{\boldsymbol{2}} = \lambda\boldsymbol{V}_1 + \boldsymbol{X}\boldsymbol{K}_{\boldsymbol{X}}\boldsymbol{X}' + \boldsymbol{V}_1\boldsymbol{X}_\perp\boldsymbol{K}_{\boldsymbol{X}_\perp}\boldsymbol{X}_\perp{}'\boldsymbol{V}_1 \tag{3.10}$$

where $\lambda \neq 0$, $\boldsymbol{X}_\perp$ is a matrix orthogonal to $\boldsymbol{X}$ so that $(\boldsymbol{X} : \boldsymbol{X}_\perp)$ is full rank, and $\boldsymbol{K}_{\boldsymbol{X}}$ and $\boldsymbol{K}_{\boldsymbol{X}_\perp}$ are arbitrary, then the BLUE of $\boldsymbol{\beta}$ is unchanged. Generally to preserve the covariance matrix of $\hat{\boldsymbol{\beta}}$, $\lambda = 1$. But if, for example, $\lambda = -1$, $\boldsymbol{K}_{\boldsymbol{X}} = \boldsymbol{0}$, $\boldsymbol{K}_{\boldsymbol{X}_\perp} = \boldsymbol{0}$, and $\boldsymbol{V}_1$ is positive semidefinite (but not the zero matrix), then $\boldsymbol{V}_2$ is not positive semidefinite.

Both Haslett & Puntanen [9] and Rao [14] are relevant here, because the diagonal elements of the $n \times n$ submatrix $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ are v_{ii}/π_i and its ijth element is v_{ij}/π_{ij}. Of course, from the population covariance $v_{ij}/(v_{ii}v_{jj})^{0.5} \leq 1$. However, in general $1/\pi_i \ll 1/\pi_{ij}$, because except perhaps for clustered designs, joint inclusion probabilities usually have the property that $\pi_{ij} \approx \pi_i\pi_j$. So diagonal elements of $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ can be much smaller than its off-diagonal elements, and consequently $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ may have at least some negative eigenvalues. The central problem has an analogue in the possibility of negative variance estimates for the Horvitz-Thompson estimator of a mean or total. The preceding material in this paper, which was also outlined in Haslett [7], forms the foundation for the following sections.

4 The augmented weighted sample superpopulation variance and adjustments to make it positive semidefinite

Recall from (3.3) that the augmented weighted sample superpopulation variance $\boldsymbol{V}_{\boldsymbol{e}_P,s}$ is defined by $\boldsymbol{V}_{\boldsymbol{e}_P,s} = \boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}^{\odot-}_{\boldsymbol{P}}$, where all three component matrices are $N \times N$, $\boldsymbol{V}_{\boldsymbol{e}_P}$ is positive semidefinite and known or estimable at least for the relevant $n \times n$ sampled submatrix, $\boldsymbol{\Pi}_{\boldsymbol{P}}$ and hence $\boldsymbol{\Pi}^{\odot-}_{\boldsymbol{P}}$ is positive but not necessarily positive semidefinite, and $\boldsymbol{\chi}_P$ is a non-negative matrix which is dependent on the sample s and contains $N^2 - n^2$ zeros plus n^2 ones. For $\boldsymbol{\chi}_P$, n of these ones are on the diagonal (corresponding to the sampled elements) and all of them can be consolidated into an $n \times n$ submatrix after suitable permutation of its rows and corresponding columns.

Now any square matrix $\boldsymbol{V}$ is positive semidefinite if and only if for conformable $\boldsymbol{x}$, $\boldsymbol{x}'\boldsymbol{V}\boldsymbol{x} \geq 0$, for all $\boldsymbol{x}$. So after any choice of suitable permutation of the rows and columns of $\boldsymbol{V}$, and by choosing conformable $\boldsymbol{x}_0 = (\boldsymbol{x}_1{}', 0, \ldots, 0)'$, then for the submatrix $\boldsymbol{V}_{11}$ corresponding to $\boldsymbol{x}_1$, $\boldsymbol{x}_1{}'\boldsymbol{V}_{11}\boldsymbol{x}_1 \geq 0$ and so is positive semidefinite.

Nevertheless, $\boldsymbol{V}_{\boldsymbol{e}_P,s}$ not being positive semidefinite is an undesirable property. To avoid such complications entirely, if the intention is to fit linear models, then the sample should if possible be chosen to meet the condition that $\boldsymbol{V}_{\boldsymbol{e}_P,s}$ is positive semidefinite, so that, for any sample s, the $n \times n$ submatrix $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ must also be positive semidefinite. This is not always possible however, in which case some adjustment may be necessary at analysis rather than design stage for estimation of linear model parameters. Two methods are discussed in Haslett [7] The first uses (3.10) and attempts to create a matrix which is positive semidefinite with the property that BLUEs are unchanged. The second uses the Cauchy interlace theorem to provide necessary but not sufficient conditions for checking, and if required reconstructing, a covariance matrix via (3.10) that meets the necessary conditions. In Huang et al. [11] there is an alternative, finding the nearest positive semidefinite matrix by minimising the (squared) Frobenius norm trace($\boldsymbol{A}\boldsymbol{A}'$) where $\boldsymbol{A}$ is the difference between the matrix in question (here $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$) and the set of positive semidefinite matrices. In essence the method when applied here would find the eigenvectors and eigenvalues of $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$, then retain the eigenvectors but replace any negative eigenvalues by zero. However the BLUE property would not necessarily be, and in fact is highly unlikely to be, maintained. For linear models though, it is perhaps not the positive definiteness of $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ that is as important as the positive semidefiniteness of $\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{X}$ since positive semidefiniteness of $\boldsymbol{X}'\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n}\boldsymbol{X}$ is sufficient to ensure that estimates of the covariance matrix of estimated regression parameters of $\boldsymbol{\beta}$ will be non-negative.

The following theorem provides a useful starting point.

Theorem 4.1. *Consider full rank but not necessarily positive semidefinite $\boldsymbol{V}$. Let $\{\lambda_i : i = 1, 2, \ldots, n\}$ be the eigenvalues of $\boldsymbol{V}$, where $\{\lambda_i : i = 1, 2, \ldots, q\}$ are all positive and $\{\lambda_i : i = q+1, q+2, \ldots, n\}$ are all negative. Let $\boldsymbol{\lambda} = \mathrm{diag}\{\lambda_i\}$, and $\boldsymbol{\lambda}_{+0}$ and $\boldsymbol{\lambda}_{-0}$ be diagonal matrices with non-zero entries of $\{\lambda_i : i = 1, 2, \ldots, q\}$ and $\{\lambda_i : i = q+1, q+2, \ldots, n\}$ respectively. Let the eigen-decomposition of*

$$\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\lambda}\boldsymbol{U}' = \begin{pmatrix} \boldsymbol{U}_{+0} & \boldsymbol{U}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_{+0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\lambda}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}'_{+0} \\ \boldsymbol{U}'_{-0} \end{pmatrix}$$

so that, since no diagonal elements of $\boldsymbol{\lambda}$ are zero,

$$\boldsymbol{V}^{-1} = \boldsymbol{U}\boldsymbol{\lambda}\boldsymbol{U}' = \begin{pmatrix} \boldsymbol{U}_{+0} & \boldsymbol{U}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}^{-1}_{+0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\lambda}^{-1}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}'_{+0} \\ \boldsymbol{U}'_{-0} \end{pmatrix}.$$

Let $\mathscr{C}$ denote column space. Then a sufficient condition for conformable $\boldsymbol{X}_0$, that $\boldsymbol{X}'_0\boldsymbol{V}^{-1}\boldsymbol{X}_0$ be positive definite, is that $\mathscr{C}(\boldsymbol{X}_0) \subseteq \mathscr{C}(\boldsymbol{V}_{+0})$.

Proof. If $\mathscr{C}(\boldsymbol{X}_0) \subseteq \mathscr{C}(\boldsymbol{V}_{+0})$ then

$$\begin{aligned}
\boldsymbol{X}'_0\boldsymbol{V}^{-1}\boldsymbol{X}_0 &= \boldsymbol{X}'_0\boldsymbol{U}\lambda^{-1}\boldsymbol{U}'\boldsymbol{X}_0 \\
&= \boldsymbol{X}'_0 \begin{pmatrix} \boldsymbol{U}_{+0} & \boldsymbol{U}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}^{-1}_{+0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\lambda}^{-1}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}'_{+0} \\ \boldsymbol{U}'_{-0} \end{pmatrix} \boldsymbol{X}_0 \\
&= \begin{pmatrix} \boldsymbol{X}'_0\boldsymbol{U}_{+0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}^{-1}_{+0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\lambda}^{-1}_{-0} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}'_{+0}\boldsymbol{X}_0 \\ \boldsymbol{0}' \end{pmatrix} \\
&= (\boldsymbol{U}'_{+0}\boldsymbol{X}_0)'\boldsymbol{\lambda}^{-1}_{+0}(\boldsymbol{U}'_{+0}\boldsymbol{X}_0).
\end{aligned}$$

which is positive definite because all the diagonal elements of $\boldsymbol{\lambda}^{-1}_{+0}$ are positive by construction.

Because many survey designs are clustered and units with clusters are correlated but not otherwise, a very common structure for $\boldsymbol{V}_{\boldsymbol{e}_P,s}$, the population error covariance, is block diagonal with blocks of size determined by the size of each cluster and all of the form $\sigma^2\{(1-\rho)\boldsymbol{I} + \rho\boldsymbol{1}\boldsymbol{1}'\}$, where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{1}$ is a vector

of ones; ρ is a positive intra-cluster correlation, and σ^2 is a scaling constant, each having the same value across clusters. Sampling from $\boldsymbol{V}_{\boldsymbol{e}_P,s}$ $(N \times N)$ then implies $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ $(n \times n)$ has a similar structure, but with reduced size because only some clusters are included in the sample (and, for a two stage cluster sample, only some units in those sampled clusters are included). Then the eigenvectors of $\boldsymbol{V}_{\boldsymbol{e}_P}$ include $\mathbf{1}$ and vectors of the form $(0, 0, \ldots, 0, 1, 1, \ldots, 1, 0, \ldots, 0)'$ where the subvector of 1's corresponds to units from a given cluster. A similar structure exists in $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ with changed dimensions because then the non-zero elements in a re-dimensioned vector $(0, 0, \ldots, 0, 1, 1, \ldots, 1, 0, \ldots, 0)'$ correspond to units sampled from a given sampled cluster. So when only the overall sample mean or the sample mean of any subsample consisting of sampled units in one or more clusters are required, the condition $\mathscr{C}(\boldsymbol{X}_0) \subseteq \mathscr{C}(\boldsymbol{V}_{+0})$ will be met, and $\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{e}_P,s,n}\boldsymbol{X}$ where $\boldsymbol{X} = \mathbf{1}$ is $n \times 1$ and defined in (2.1) will be positive definite even if $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ is not.

These results a consequence of the following lemma.

Lemma 4.1. *$\boldsymbol{V} = a\boldsymbol{I} + b\mathbf{1}\mathbf{1}'$ and $\mathbf{1}$ share the eigenvector $\mathbf{1}$.*

Proof. Let $\boldsymbol{V}$ be $n_c \times n_c$ and $\mathbf{1}$ be $n_c \times 1$. Clearly $\mathbf{1}$ is an eigenvector of itself with eigenvalue 1. Further, $\mathbf{1}$ is an eigenvector of $\boldsymbol{V}$ with eigenvalue $a + n_c b$, since $(a\boldsymbol{I} + b\mathbf{1}\mathbf{1}')\mathbf{1} = (a + n_c b)\mathbf{1}$ via $\mathbf{1}'\mathbf{1} = n_c$.

Aggregations of clusters or sampled clusters will give a block diagonal covariance structure corresponding to population or sample submeans, or the population or sample overall mean, respectively, and (via the result for the clusters) will yield the required positive semidefiniteness. In the scalar case this leads to a positive estimate of variance for both sample submeans and the overall sample mean.

5 Sampling with replacement

One implicit assumption in the preceding material is that the n units sampled from a complex survey design are all distinct, or if not then when a unit is redrawn it is assumed not to have a correlation of one with any previous or consequent draw of the same unit. However, much of the previous discussion can be extended to varying probability sampling with replacement, in which sampling of units may be with unequal probability, but any sampled unit may be resampled. In this case the number of unique units in the sample n_u may be and often is less than n, the sample size. The difference $(n - n_u)$ represents the number of units that are replicates. For sampled unit k, let n_k equal the number of times the unit is sampled. Of course, for the unsampled units, $n_k = 0$. A commonly used sampling scheme that uses varying probability is probability proportional to size (pps) sampling, where the variable of interest is positively correlated with some measure of unit size known before data collection. This measure of size is used directly at survey design stage as the basis for setting the probability that each unit is included in the sample. Sampling without replacement is usually but not necessarily more efficient (see Gabler [5] for example). Much depends on the joint inclusion probabilities, as illustrated in Rao [15]. The probability any unsampled unit is sampled also changes with each draw. Achieving an overall pps sample without replacement (ppswor) is consequently not as straightforward as sampling pps with replacement (ppswr). This is the principal reason ppswr is a common sampling scheme, especially in third world applications. Formulae for estimating a mean, its variance and estimated variance for ppswr are well known. But when a linear model is fitted to any with replacement sample, the complication is that, for any sample where one or more units are replicated in the sample, the covariance structure for all n sampled units is singular.

There are two options. The first is based on the unique units sampled only, ignoring any replicates. This is a prevalence estimator, since no matter how many times it occurs in a particular sample a unit is counted only once in the estimator. Here the sample size is random so an issue is determining design unbiasedness taken over all possible samples. The second is an incidence estimator, which incorporates into the estimator the number of times a unit is sampled in any particular sample. Here the complication is that all occurrences in the sample of a given unit have a correlation of one with each other so, for any sample with replicated units,

matrix inversion for linear models involves generalized inverses. We now consider each of these situations in turn.

5.1 The prevalence method

Suppose sampling is varying probability with replacement and that the probability that unit i is selected at each draw is p_i . Then for all $i = 1, 2, \ldots, N$, the inclusion probability is

$$\begin{aligned}\pi_i &= prob(i \in s)\\ &= 1 - prob(i \notin s)\\ &= 1 - (1 - p_i)^n.\end{aligned}$$

The joint inclusion probability for units i and j is

$$\begin{aligned}\pi_{ij} &= prob(i \in s, j \in s)\\ &= 1 - prob(i \notin s) - prob(j \notin s) + prob(i \notin s, j \notin s)\\ &= 1 - (1 - p_i)^n - (1 - p_j)^n + (1 - p_i - p_j)^n.\end{aligned}$$

Further,

$$prob(i \in s, r \text{ times}) = {}_nC_r p_i^r (1 - p_i)^{n-r}$$

where ${}_nC_r = n!/\{r!(n-r)!\}$ is the number of ways of choosing r items from n.

The joint probability that unit i appears exactly r times, and the unit j appears exactly t times is given via the multinomial expansion,

$$prob(i \in s, r \text{ times}; j \in s, t \text{ times}) = \{n!/(r!t!(n-r-t)!\} p_i^r p_j^t (1 - p_i - p_j)^{n-r-t}.$$

The maximum number of times unit i can be selected in a sample of size n is n, and

$$\sum_{r=1}^{n} {}_nC_r p_i^r (1 - p_i)^{n-r} = \sum_{r=0}^{n} {}_nC_r p_i^r (1 - p_i)^{n-r} - (1 - p_i)^n = 1 - (1 - p_i)^n.$$

Note as an aside, that

$$\sum_{r=1}^{n} {}_nC_r p_i^r (1 - p_i)^{n-r} = 1 - (1 - p_i)^n = 1 - \sum_{r=0}^{n} {}_nC_r (-p_i)^r (1)^{n-r} = \sum_{r=1}^{n} {}_nC_r (-1)^{r-1} {p_i}^r$$

provides a connection between a series expansion with entirely positive terms and one with terms of alternating sign.

Lemma 5.1. *Godambe [6] The expected number of unique units in a with replacement sample with equal probabilities of selection at each draw for each unit $i = 1, 2, \ldots, , N$ but possibly varying probabilities of selection for different units, is the sum of the inclusion probabilities of all N population units, i.e.,*

$$\mathrm{E}(n_u) = \sum_{i=1}^{N} \pi_i. \tag{5.1}$$

Proof. If we sum the probability of each unit in the sample appearing $1, 2, \ldots, n$ times to create a sample of size n_u (the number of unique units) from it, this is equivalent to only counting unit i once no matter how many times more than zero it appears in the sample. Then the probability that unit i appears in the reduced set of size n_u is the sum of the probabilities it appears in the sample $\{k = 1, 2, \ldots, n\}$ times, i.e.,

$$\sum_{r=1}^{n} {}_nC_r p_i^r (1 - p_i)^{n-r} = 1 - (1 - p_i)^n = \pi_i.$$

It follows, since probability that unit i appears in the reduced set does not depend on the joint inclusion probabilities, that even if there were clustering in the sampling scheme, taken over all samples the expected number of unique units in the sample is the sum of all N inclusion probabilities, namely $\{\pi_i : i = 1, 2, \ldots, N\}$, so that $\mathrm{E}(n_u) = \sum_{i=1}^{N} \pi_i$.

Lemma 5.2. *The sum of the inclusion probabilities for all N units for a with replacement sample with equal probabilities of selection at each draw for each unit $i = 1, 2, \ldots, N$ but possibly varying probabilities of selection for different units is less than the sample size n , i.e., $\sum_{i=1}^{N} \pi_i < n$.*

Proof: Using the previous transition from a series with all terms positive to one where terms have alternating signs. We have

$$\begin{aligned}
\sum_{i=1}^{N} \pi_i &= \sum_{i=1}^{N} \{1 - (1 - p_i)^n\} \\
&= \sum_{i=1}^{N} \sum_{r=1}^{n} {}_nC_r (-1)^{r-1} p_i^r \\
&= \sum_{r=1}^{n} {}_nC_r \sum_{i=1}^{N} (-1)^{r-1} p_i^r \\
&= \sum_{r=1}^{n} {}_nC_r (-1)^{r-1} \sum_{i=1}^{N} p_i^r \\
&< \sum_{r=1}^{n} {}_nC_r (-1)^{r-1} \text{since} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{p}_\mathrm{i}^\mathrm{r} < \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{p}_\mathrm{i} = 1 \\
&=_n C_0 - \sum_{r=0}^{n} {}_nC_r (-1)^r (1)^{n-r} \\
&= n - (-1 + 1)^n \\
&= n.
\end{aligned}$$

So $\sum_{i=1}^{N} \pi_i < n$ for varying probability sampling with replacement in general and for ppswr in particular. More intuitively, $\sum_{i=1}^{N} \pi_i = \mathrm{E}(n_u)$ from Lemma 5.1 and sampling is with replacement so that $\mathrm{E}(n_u) < n$, and it follows that $\sum_{i=1}^{N} \pi_i < n$.

The covariance matrix of the unique sample elements from a with replacement sample is $N \times N$ with the same structure as previously namely (3.3), $\boldsymbol{V}_{\boldsymbol{e}_P,S} = \boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$. Here the diagonal elements of $\boldsymbol{\chi}_P$ indicate those units i in the population P that are selected in the sample at least once, so that $\boldsymbol{\chi}_P$ after appropriate reordering contains an $(n_u \times n_u)$ submatrix for which all elements are equal to one, with all other elements zero. Further, as discussed in Haslett [8], the elements in $\boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ are the elementwise inverses of $\pi_i = 1 - (1 - p_i)^n$ on the diagonal, and $\pi_{ij} = 1 - (1 - p_i)^n - (1 - p_j)^n + (1 - p_i - p_j)^n$ in the (i, j)th off-diagonal position, and $\boldsymbol{V}_{\boldsymbol{e}_P}$ is the population covariance matrix structure as before. Thus, from (3.8) and (3.9) with $\boldsymbol{X}_{n_u}$ $(n_u \times p)$, defined to be that part of $\boldsymbol{X}$ $(n \times p)$ that corresponds to the unique sample elements, as is $\boldsymbol{Y}_{n_u}$, the approximate BLUE based on the sampled elements is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'_{n_u} \boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u} \boldsymbol{X}_{n_u})^{-1} \boldsymbol{X}'_{n_u} \boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u} \boldsymbol{Y}_{n_u} \tag{5.2}$$

with

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = \{(\boldsymbol{X}'_{n_u} \boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u} \boldsymbol{X}_{n_u})^{-1} \boldsymbol{X}'_{n_u} \boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u}\} \widehat{\boldsymbol{V}_{\boldsymbol{e}}} \{\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u} \boldsymbol{X}_{n_u} (\boldsymbol{X}'_{n_u} \boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u} \boldsymbol{X}_{n_u})^{-1}\} \tag{5.3}$$

As in Section 2, $\boldsymbol{V}(\boldsymbol{e}_{n_u}) = E(\boldsymbol{e}_{n_u}\boldsymbol{e}'_{n_u})$, with $\boldsymbol{e}_{n_u}$ now denoting that part of $\boldsymbol{e}_P$ that corresponds to the n_u unique sampled elements, and $\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u}$ being $(n_u \times n_u)$. The size of the various matrices and vectors varies from sample to sample with expected number of unique elements $\mathrm{E}(n_u) = \sum_{i=1}^{n} \pi_i$. As required by (3.4) and Theorem 3.1, despite this contraction in average dimension,

$$\mathrm{E}(\boldsymbol{V}_{\boldsymbol{e}_P,s}) = \mathrm{E}(\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}^{\odot -}_{\boldsymbol{P}}) = \boldsymbol{V}_{\boldsymbol{e}_P}$$

$$\mathrm{E}(\boldsymbol{X}'_{n_u}\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u}\boldsymbol{X}_{n_u}) = (\boldsymbol{X}'_P\boldsymbol{V}^{-1}_{\boldsymbol{e}_P}\boldsymbol{X}_P) \tag{5.4}$$

$$\mathrm{E}(\boldsymbol{X}'_{n_u}\boldsymbol{V}^{-1}_{\boldsymbol{e}_P,s,n_u}\boldsymbol{Y}_{n_u}) = (\boldsymbol{X}'_P\boldsymbol{V}^{-1}_{\boldsymbol{e}_P}\boldsymbol{Y}_P)$$

Given $\boldsymbol{X}_P$ is fixed under the superpopulation model, i.e., for each $i = 1, 2, \ldots, N$ the corresponding row of $\boldsymbol{X}_P$ is fixed, then the three equations in (5.4) also hold for the joint design-superpopulation expectation.

5.2 The incidence method

Suppose again that sampling is varying probability with replacement and that the probability that unit i is selected at each draw is p_i. Inclusion and joint inclusion probabilities, and probability unit i is drawn r times, and the probability that unit i is drawn r times and unit j is drawn t times have been given in Subsection 5.1.

Also, since $prob(i \in s, r \text{ times}) = {}_nC_r p_i^r (1-p_i)^{n-r}$, the expected number of times unit i appears in the sample is

$$\sum_{r=1}^{n} r\, {}_nC_r p_i^r (1-p_i)^{n-r} = np_i.$$

This is so since $(a+b)^n = \sum_{r=0}^{n} {}_nC_r a^r b^{n-r}$, hence the partial derivative is

$$\partial(a+b)^n/\partial a = \sum_{r=0}^{n} r\, {}_nC_r a^{r-1} b^{n-r}.$$

Thus

$$a(\partial(a+b)^n/\partial a) = an(a+b)^{n-1} = \sum_{r=0}^{n} r\, {}_nC_r a^r b^{n-r}$$

and the expected number of times unit i appears in the sample is

$$\sum_{r=1}^{n} r\, {}_nC_r p_i^r (1-p_i)^{n-r} = np_i\{p_i + (1-p_i)\}^n - \{r\, {}_nC_r p_i^r (1-p_i)^{n-r}\}_{r=0} = np_i.$$

5.2.1 Generalized inverses for with replacement sampling

For definiteness in what follows, define the Moore-Penrose (M-P) inverse $\boldsymbol{A}^+$ of a real matrix $\boldsymbol{A}$ by the following four properties:

1. $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$
2. $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$
3. $\boldsymbol{A}^+\boldsymbol{A} = (\boldsymbol{A}^+\boldsymbol{A})'$
4. $\boldsymbol{A}\boldsymbol{A}^+ = (\boldsymbol{A}\boldsymbol{A}^+)'$.

Consider a sample of size n sampled with replacement, where only unit j is resampled, so that $n_u = n-1$. To represent the sampling mechanism, order the n_u unique elements first, followed by the second occurrence of unit j.

Let

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{i}_j \\ \boldsymbol{i}_j' & \boldsymbol{i}_j'\boldsymbol{i}_j \end{pmatrix}$$

where

$$\boldsymbol{i}_j = (\ 0 \quad 0 \quad \ldots \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0)'$$

has one in the jth position with all other elements zero. Let

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Sigma} & 0 \\ 0' & 0 \end{pmatrix}$$

where $\boldsymbol{\Sigma}_0$ is the covariance structure for the n_u unique elements. Then $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}' = \boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$ is the singular covariance structure for all n sampled units.

Note that

$$\boldsymbol{i}_j'\boldsymbol{i}_j = 1;\ \boldsymbol{i}_j\boldsymbol{i}_j' = \mathrm{diag}(\boldsymbol{i}_j);\ \boldsymbol{i}_j'\boldsymbol{i}_j\boldsymbol{i}_j' = \boldsymbol{i}_j';\ \boldsymbol{i}_j\boldsymbol{i}_j'\boldsymbol{i}_j = \boldsymbol{i}_j;\ \text{and } \boldsymbol{I} + \boldsymbol{i}_j\boldsymbol{i}_j' = \mathrm{diag}(n_k)$$

where n_k is the number of occurrences of sampled unit k.

Then

$$\boldsymbol{T}^* = \begin{pmatrix} I + \boldsymbol{i}_j\boldsymbol{i}_j' & -\boldsymbol{i}_j \\ -\boldsymbol{i}_j' & \boldsymbol{i}_j'\boldsymbol{i}_j \end{pmatrix}$$

is a (1-) g-inverse of $\boldsymbol{T}$. Further, assuming $\boldsymbol{\Sigma}$ is full rank, let

$$\boldsymbol{\Sigma}_0^+ = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & 0 \\ 0' & 0 \end{pmatrix}$$

be the M-P inverse of $\boldsymbol{\Sigma}$. Then $\boldsymbol{T}^*\boldsymbol{\Sigma}_0^+\boldsymbol{T}^*$ is a (1,2-) g-inverse of $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$.

This structure can be extended to include units sampled more than two times. In such cases, the (2,2) block of $\boldsymbol{T}$ is itself block diagonal with diagonal blocks that are all ones, of size $(n_k - 1)x(n_k - 1)$. So if $n_k > 2$ for any $k = 1, 2, \ldots, n_u$, the (2,2) block of $\boldsymbol{T}$ is not always full rank.

Consequently, the conditions of Puntanen et al. [13, p.294], which provide an explicit Moore-Penrose inverse for a block matrix where the (1,1) and (2,2) blocks are both invertible, are not met. Further, although Jerković & Malešević [12] provides conditions for various types of g-inverse, which include (1-) and (1,2-) g-inverses, their results provide equivalent conditions rather than a method of construction. The question then is how to find suitable g-inverses of $\boldsymbol{T}$ and $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$ when there may be replicates ($n_k > 1$) for any number of the n_u sampled units.

Consider

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{L}' \\ \boldsymbol{L} & \boldsymbol{L}\boldsymbol{L}' \end{pmatrix}.$$

For with replacement sampling, $\boldsymbol{L}$ has a particular structure with all elements zero except that the $(n_k - 1)$ rows for $n_k > 1$ corresponding to replicates (i.e., repetitions) of the kth unique unit have a one in column k.

Theorem 5.1. *Let*

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{L}' \\ \boldsymbol{L} & \boldsymbol{L}\boldsymbol{L}' \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{0} \end{pmatrix}$$

where the blocks of $\boldsymbol{T}$ and $\boldsymbol{\Sigma}_0$ conform. For simplicity, let $\boldsymbol{\Sigma}$ be full rank. Let

$$\boldsymbol{\Sigma}_0^+ = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{0} \end{pmatrix}$$

be the M-P inverse of $\boldsymbol{\Sigma}_0$. *Then given* $\boldsymbol{L}'$ *is a (1,2-) g-inverse of* $\boldsymbol{L}$, *a (1-) g-inverse of* $\boldsymbol{T}$ *is*

$$\boldsymbol{T}^* = \begin{pmatrix} \boldsymbol{I} + \boldsymbol{L}'\boldsymbol{L} & -\boldsymbol{L}' \\ -\boldsymbol{L} & \boldsymbol{L}\boldsymbol{L}' \end{pmatrix}$$

and for any $\boldsymbol{L}$, *a (1,2-) g-inverse of* $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$ *is* $\boldsymbol{T}^*\boldsymbol{\Sigma}_0^+\boldsymbol{T}^*$.

Proof: By expansion.

Note that more general forms of $\boldsymbol{T}^*$ are possible, and there exists a complete class of such (1-) g-inverses which have the same structure as in Theorem 5.1, but for which each of the blocks is scaled by a different matrix. However none of the members of this class satisfy conditions 3- or 4- for M-P inverses.

Corollary 5.1. *When* $\boldsymbol{\Sigma}_0$ *is the covariance structure for the* n_u *unique elements from a with replacement sample with varying probabilities, and* $\boldsymbol{L}$ *has all elements zero except that all* $(n_k - 1)$ *rows for* $n_k > 1$ *corresponding to replicates (i.e., repetitions) of the kth unique sampled unit have a one in column k, then Theorem 5.1 provides a (1-) g-inverse of* $\boldsymbol{T}$ *and a (1,2-) g-inverse of* $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$.

Note that, for with replacement sampling, $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$ is the covariance structure for all n sampled elements including repetitions, and $\boldsymbol{T}^*\boldsymbol{\Sigma}_0^+\boldsymbol{T}^*$ is a (1,2-) g-inverse of $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$.

5.2.2 Linear models using incidence for with replacement sampling

Because $\boldsymbol{T}^*\boldsymbol{\Sigma}_0^+\boldsymbol{T}^*$ is a g-inverse of $\boldsymbol{T}\boldsymbol{\Sigma}_0\boldsymbol{T}$, the possibly singular covariance structure for all n sampled elements including repetitions, we now have the matrices needed to specify and fit a linear model to sample survey data from a varying probability with replacement sample using all n sampled units and allowing for the correlation between any unit that is resampled. Specifically from (3.8) and (3.9) for full rank $\boldsymbol{X}'\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\boldsymbol{X}$ one parameterisation of the approximate BLUE based on all n sampled elements is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\boldsymbol{Y} \tag{5.5}$$

with estimated covariance

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = \{(\boldsymbol{X}'\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\}\widehat{\boldsymbol{V}_{\boldsymbol{e}}}\{\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^-_{\boldsymbol{e}_P,s,n}\boldsymbol{X})^{-1}\} \tag{5.6}$$

where, as in Section 2 and Subsection 3.2, $\boldsymbol{V}(\boldsymbol{e}) = \boldsymbol{V}_{\boldsymbol{e}}$. and $\boldsymbol{e}$ denotes that part of $\boldsymbol{e}_P$ that corresponds to the n sampled elements.

Here $\boldsymbol{V}_{\boldsymbol{e}_P,s,n}$ is an $n \times n$ submatrix derived from an extended $nN \times nN$ form of the augmented weighted sample superpopulation variance, $\boldsymbol{V}_{\boldsymbol{e}_P,s} = \boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ This extended form considers separately the first through nth time that each of the N population units may be sampled, and includes both the n_u sampled elements and $(N - n_u)$ unsampled elements. Each of the N population units is included n times, even though only n_u distinct elements are sampled in a given sample of size n.

Denoting the Kronecker product of matrices by $\otimes$, the extended $nN \times nN$ form of $\boldsymbol{\chi}_P \odot \boldsymbol{V}_{\boldsymbol{e}_P} \odot \boldsymbol{\Pi}_{\boldsymbol{P}}^{\odot -}$ is

$$\{\boldsymbol{\chi}_{P_{Nn}} \odot (\boldsymbol{1}\boldsymbol{1}' \otimes \boldsymbol{V}_{\boldsymbol{e}_P}) \odot \boldsymbol{\Pi}_{P_{Nn}}^{\odot -}\}$$

which being $nN \times nN$ allows that each population unit $i = 1, 2, \ldots, N$ may occur in a sample $k = 1, 2, \ldots, n$ times. Of course for any given sample s, only an $n \times n$ submatrix based on the n_u unique sampled elements contains non-zero rows and columns, because only an $n \times n$ submatrix of the $nN \times nN$ matrix $\boldsymbol{\chi}_{P_{Nn}}$ contains non-zero elements. $\boldsymbol{\Pi}_{P_{Nn}}^{\odot -}$ is also $nN \times nN$, with elements equal to the inverse of the relevant terms in a binomial expansion (for diagonal elements) or a multinomial expansion (for off-diagonal elements). See the initial equations in Section 5.2.

Then

$$\boldsymbol{V}_{\boldsymbol{e}_P,S} = (\mathbf{1}_n \otimes \boldsymbol{I}_N)'\{\boldsymbol{\chi}_{P_{Nn}} \odot (\mathbf{1}\mathbf{1}' \otimes \boldsymbol{V}_{\boldsymbol{e}_P}) \odot \boldsymbol{\Pi}_{P_{Nn}}^{\odot -}\}(\mathbf{1}_n \otimes \boldsymbol{I}_N) \tag{5.7}$$

which combines instances of multiple sampling, is a design unbiased estimator of $\boldsymbol{V}_{\boldsymbol{e}_P}$. Of course, for a particular sample only $n \times n$ submatrix is involved, because all other $(Nn)^2 - n^2 = n^2(N^2 - 1)$ entries of $\boldsymbol{V}_{\boldsymbol{e}_P,S}$ are zero.

6 Conclusions

When survey estimation is design based and includes weights, via inclusion (or selection) and joint inclusion (or selection) probabilities, or functions of them (for example, a non-response adjustment), the positive semidefiniteness of estimated covariance structure for the error in a linear model constructed from survey data cannot be guaranteed, except if all joint inclusion probabilities, or equivalently covariances between population elements are ignored. For particular types of covariance structure often used in linear models for survey data however, where the covariance matrix is block diagonal with common correlation and scale such as often used for cluster sampling and its variants, it is possible to ensure positive definiteness for the estimated covariance matrix of parameter estimates. The methods previously used in Haslett [7] for sampling without replacement can be extended to with replacement sampling schemes even with varying probabilities for sampling units and clustered sampling. The principal approaches are to consider only the unique units sampled, or to work with all units sampled and possibly singular covariance structures.

Although Section 5 has focused on prevalence and incidence type estimators for design based and for joint design and model-based inference given $\boldsymbol{X}_P$, the various generalized inverses developed there are also applicable to model-based sampling when sampling is with replacement. One simplification then is that the positive semidefiniteness considerations due to selection probabilities that are necessary for joint design and model approach are not required for model-based inference because the selection mechanism is instead incorporated into the auxiliary variables in the model rather than using inclusion or selection probabilities. The sample covariance for the unique sampled units is consequently positive semidefinite because the issues induced by the matrix of elementwise inclusion and joint inclusion probabilities is not relevant. The complication that the sample covariance matrices will be positive semidefinite but not positive definite will remain when any unit is sampled more than once, but the generalized inverses in Section 5 can be used to circumvent this problem and (as for the design based and the joint and design model based framework) can provide sample based BLUEs.

References

[1] Chambers, R. & Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series, Oxford University Press.

[2] Chambers, R. & Skinner, C.J. (2003). *Analysis of Survey Data*, Wiley.

[3] Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition, Wiley.

[4] Fuller, W. A. (2009). *Sampling Statistics*, Wiley.

[5] Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement, *Biometrika*, 71(1), 171-175.

[6] Godambe, V. P. (1955). A unified theory of sampling from finite populations, *Journal of the Royal Statistical Society* B, 17, 269-278.

[7] Haslett, S. (2016). Positive semidefiniteness of estimated covariance matrices in linear models for sample survey data, *Special Matrices*, 4, 218-224.

[8] Haslett, S. (1985). The linear non-homogeneous estimator in sample surveys, *Sankhyä* B, 47, 101-117.

[9] Haslett, S. & Puntanen, S. (2010). Equality of the BLUEs and/or BLUPs under two linear models using stochastic restrictions, *Statistical Papers*, 51, 465-475.

[10] Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47, 663-685.

[11] Huang, C., Farewell, D. & Pan, J. (2017). A calibration method for non-positive definite covariance matrix in multivariate analysis, *Journal of Multivariate Analysis*, 157, 45-52.
[12] Jerković, V. M. & Malešević, B. (2014). Block representations of generalized inverses of matrices, *Symposium on Mathematics and Its Applications*, Faculty of Mathematics, University of Belgrade, Vol V(1), 10 pages. https://arxiv.org/ftp/arxiv/papers/1509/1509.03458.pdf
[13] Puntanen, S., Styan, G.P.H. & Isotalo, J. (2011). *Matrix Tricks for Linear Statistical Models*, Springer.
[14] Rao, C. R. (1968). A note on a previous lemma in the theory of least squares and some further results, *Sankhyä* A, 30, 259-266.
[15] Rao, J. N. K. (1963). On two systems of unequal probability sampling without replacement, *Annals of the Institute of Statistical Mathematics*, 15, 67-72.
[16] Royal, R. M. & Cumberland, W. G. (1978). Variance estimation in finite population sampling, *Journal of the American Statistical Association*, 73, 351-358.
[17] Särndal, C-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer.
[18] Skinner, C.J., Holt, D. & Smith, T.M.F. (1989). *Analysis of Complex Surveys*, Wiley.