Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 2991–3004
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Predicting random effects with an expanded finite population mixed model

Edward J. Stanek III[a,*], Julio M. Singer[b]

[a] *Department of Public Health, 401 Arnold House, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304, USA*
[b] *Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil*

Received 31 July 2007; received in revised form 6 November 2007; accepted 20 November 2007
Available online 31 December 2007

Abstract

Prediction of random effects is an important problem with expanding applications. In the simplest context, the problem corresponds to prediction of the latent value (the mean) of a realized cluster selected via two-stage sampling. Recently, Stanek and Singer [Predicting random effects from finite population clustered samples with response error. J. Amer. Statist. Assoc. 99, 119–130] developed best linear unbiased predictors (BLUP) under a finite population mixed model that outperform BLUPs from mixed models and superpopulation models. Their setup, however, does not allow for unequally sized clusters. To overcome this drawback, we consider an expanded finite population mixed model based on a larger set of random variables that span a higher dimensional space than those typically applied to such problems. We show that BLUPs for linear combinations of the realized cluster means derived under such a model have considerably smaller mean squared error (MSE) than those obtained from mixed models, superpopulation models, and finite population mixed models. We motivate our general approach by an example developed for two-stage cluster sampling and show that it faithfully captures the stochastic aspects of sampling in the problem. We also consider simulation studies to illustrate the increased accuracy of the BLUP obtained under the expanded finite population mixed model.

Keywords: Superpopulation; Best linear unbiased predictor; Random permutation; Optimal estimation; Design-based inference; Mixed models

1. Introduction

Optimal estimation of average costs for hospitals that typically vary in size is an important practical problem because of the impact in health care economics, and patient choice of hospital care (see http://www.healthgrades.com, for example). In many cases, this is based on information obtained from patients (units) in hospitals (clusters) realized under a two-stage sampling scheme.

The best linear unbiased predictor (BLUP) developed under a mixed model is often offered as a solution to this problem (Searle et al., 1992). Although the mixed model accounts for unequal numbers of units in sample clusters, it does not use often available information about their sizes. The superpopulation model of Scott and Smith (1969) is an alternative that incorporates this information. Both models can be plausibly used to represent the problem of interest, but neither is formally linked to the finite population from which the two-stage sample is drawn as is the finite

* Corresponding author. Tel.: +1 413 545 3812; fax: +1 413 545 1645.
E-mail addresses: stanek@schoolph.umass.edu (E.J. Stanek III), jmsinger@ime.usp.br (J.M. Singer).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.11.012

Table 1
Population of hospital's appendectomy patients in the past year and observed data

Hospital (s)	M_s	Mean	Variance	Patient* (t)			
$s=1$ (County)	2	μ_1	σ_1^2	y_{11}	y_{12}		
$s=2$ (Central)	4	μ_2	σ_2^2	y_{21}	y_{22}	y_{23}	y_{24}
				\$2100 (Jane Blake)		\$1400 (Sam Evans)	\$2500 (Hong Yao)
$s=3$ (Mercy)	2	μ_3	σ_3^2	y_{31}	y_{32}	y_{33}	
				\$1700 (Mary Slokum)	\$1900 (Juan Marcus)		

* Names are fictitious

population mixed model recently proposed by Stanek and Singer (2004)[1] for situations where clusters are of equal size. Under this model, predictors have smaller mean squared error (MSE) than the competitors, even when the variance components are replaced by estimates as indicated in San Martino et al. (2008). We extend the approach of Stanek and Singer (2004) by developing predictors under a new expanded finite population mixed model that outperforms the competitors both in equal and unequal size two-stage cluster sampling problems.

Suppose our interest is in the average cost of appendectomies (the latent value) for each of three hospitals in the past year (Table 1), and that such costs are known (without error) for some patients in two of the hospitals. When the data are obtained from a stratified simple random sample of appendectomy patients, with hospitals as strata, the best linear unbiased estimate is the average cost for the available patients in each hospital (i.e., \$2000 for *Central*, and \$1800 for *Mercy*).

Now assume that a simple random sample of appendectomy patients is selected from each of a simple random sample of hospitals (Table 2) according to a two-stage sampling scheme. We refer to a sample hospital as a primary sampling unit (PSU) to distinguish it from a specific hospital, and to a sample patient as a secondary sampling unit (SSU) to distinguish it from a specific patient. Under the usual mixed model, the sample appendectomy cost for SSU j in PSU i is

$$Y_{ij} = \mu + B_i + E_{ij}, \tag{1}$$

where μ is the overall mean, B_i is the random effect for PSU i, and E_{ij} is a random variable corresponding to the deviation of the response of SSU j from the latent value of PSU i, namely $T_i = \mu + B_i$. The random variables B_i and E_{ij} are usually considered independent with null expected values, and variances given by σ^2 and σ_i^2, respectively. Model (1) is an example of the general linear mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{B} + \mathbf{E}, \tag{2}$$

where for the sample in Table 2, $\mathbf{X} = \mathbf{1}_r$, $\mathbf{Z} = \oplus_{i=1}^n \mathbf{1}_{m_i}$, $\boldsymbol{\alpha} = \mu$, and $\mathbf{B} = (B_1, \ldots, B_n)'$ with $\boldsymbol{\Gamma} = \sigma^2\mathbf{I}_n$, $\boldsymbol{\Sigma} = \oplus_{i=1}^n \sigma_i^2\mathbf{I}_{m_i}$, and $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Omega} = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}' + \boldsymbol{\Sigma}$ with $\mathbf{1}_a$ denoting an $a \times 1$ vector with all elements equal to one, $\mathbf{I}_a$ representing an $a \times a$ identity matrix, and $\oplus_{i=1}^n \mathbf{A}_i$ indicating a block diagonal matrix with blocks given by $\mathbf{A}_i$ (Graybill, 1983). This model has a long history (see for example Harville, 1978; Laird and Ware, 1982) and is the main topic in several recent texts such as Brown and Prescott (1999), Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Bryk and Raudenbush (2002), Diggle et al. (2002), Singer and Willett (2003), Demidenko (2004), Littell et al. (2006), and Jiang (2007). Under (1), the BLUP of the latent value for PSU i is

$$\hat{P}_i = \hat{\mu} + k_i(\bar{Y}_i - \hat{\mu}), \tag{3}$$

[1] We refer to such models as a finite population mixed models instead of random permutation models as in Stanek and Singer (2004) to avoid confusion with the homonymous, but different, model considered in Hedayat and Sinha (1991).