Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 2199–2217
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Performance of balanced two-stage empirical predictors of realized cluster latent values from finite populations: A simulation study

Silvina San Martino[a], Julio M. Singer[b,*], Edward J. Stanek III[c]

[a] *Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina*
[b] *Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil*
[c] *Department of Public Health, University of Massachusetts, Amherst, MA, USA*

Received 24 January 2007; received in revised form 16 July 2007; accepted 19 July 2007
Available online 31 July 2007

Abstract

Predictors of random effects are usually based on the popular mixed effects (*ME*) model developed under the assumption that the sample is obtained from a conceptual infinite population; such predictors are employed even when the actual population is finite. Two alternatives that incorporate the finite nature of the population are obtained from the superpopulation model proposed by Scott and Smith (1969. Estimation in multi-stage surveys. J. Amer. Statist. Assoc. 64, 830–840) or from the finite population mixed model recently proposed by Stanek and Singer (2004. Predicting random effects from finite population clustered samples with response error. J. Amer. Statist. Assoc. 99, 1119–1130). Predictors derived under the latter model with the additional assumptions that all variance components are known and that within-cluster variances are equal have smaller mean squared error (*MSE*) than the competitors based on either the *ME* or Scott and Smith's models. As population variances are rarely known, we propose method of moment estimators to obtain empirical predictors and conduct a simulation study to evaluate their performance. The results suggest that the finite population mixed model empirical predictor is more stable than its competitors since, in terms of *MSE*, it is either the best or the second best and when second best, its performance lies within acceptable limits. When both cluster and unit intra-class correlation coefficients are very high (e.g., 0.95 or more), the performance of the empirical predictors derived under the three models is similar.

Keywords: Empirical predictors; Finite population; Optimal prediction; Random permutation; Two-stage sampling; Mixed model; Superpopulation; Cluster sampling; Hierarchical models

1. Introduction

There are many instances where clustered finite populations occur naturally as in educational, public health or sociological surveys, where classrooms in schools, physician practices in hospitals or families in communities are typical examples of such clusters. In such settings, statistical inference is usually based on a multi-stage random sample selected without replacement. In addition to the sample average, three approaches may be considered to predict latent values of realized clusters (i.e., the average expected response of the units in those clusters). In each case, best linear

* Corresponding author. Tel.: +55 11 30916118; fax: +55 11 30916130.
E-mail addresses: ssanmarti@balcarce.inta.gov.ar (S. San Martino), jmsinger@ime.usp.br (J.M. Singer), stanek@schoolph.umass.edu (E.J. Stanek).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2007.07.013

unbiased predictors (BLUP) have been obtained. The most popular approach is based on the usual mixed effects (*ME*) model derived under the assumption that the sample is obtained from a conceptual infinite population and is considered in Goldberger (1962), Henderson (1984), Kackar and Harville (1984), Prasad and Rao (1990), McLean et al. (1991), Robinson (1991), Stanek et al. (1999), Moura and Holt (1999) or McCulloch and Searle (2001) among others. The second approach, suggested by Scott and Smith (1969) and extended by Bolfarine and Zacks (1992) to include response error, considers the finite nature of the population and bases the inference on a superpopulation (*SP*) model. It has limited application, in part because its performance may be affected by model miss-specification. The third approach, recently suggested by Stanek et al. (2004) in a simple random sampling setup and extended by Stanek and Singer (2004) to a balanced two-stage sampling setup with or without response error, considers a finite population mixed (*FM*) model,[1] induced by the sampling process. Since the stochastic model is developed directly from two-stage sampling from a finite population, it can be applied to a wide range of practical settings.

In each case, the BLUPs of realized cluster latent values involve predictors of the response of the unobserved units and depend on weights called shrinkage factors; these shrinkage factors are functions of population variance components and of the number of sampled clusters. For the last two models, they also depend on finite population characteristics such as cluster sizes. The predictors obtained under these three models can occasionally be quite similar, but sometimes they can differ greatly.

As an example, suppose that an educational survey is conducted in a given high-school to evaluate the ability of second graders in a certain subject by means of a test with scores ranging from 0 to 10. We assume that the student responses include response error. To account for teacher effects, a two-stage random sample is obtained from the population of second grade students assigned to classrooms (each with 30 students). Assume that a sample of 15 students is selected from each classroom in a sample of classrooms in the school. In addition to estimating the school response and variance components, suppose that there is interest in predicting the latent response for a sampled classroom. For illustration, let us assume that the between classroom variance is 1.25, the within-cluster variance is 2.00 and that the response error variance is 0.80. With these assumptions, the cluster intra-class correlation coefficient is 0.38 and the unit intra-class correlation coefficient is 0.71 (see Section 2.4 for definitions of cluster and unit intra-class correlation coefficients). Also, suppose that the school sample average is 6.75, while for the ith sample classroom, the sample average is 5.20. Based on the sample data, there are four approaches to predict the latent classroom response. First, we may use the sample classroom average that is 5.20. Alternatively, assuming that the response error model holds for all students, the latent response for the ith sample classroom is predicted to be 5.40, 5.30 and 5.90, respectively, using the *ME*, *SP* or *FM* model predictors (see Section 2.4 for details about each predictor). The 11% relative difference observed between the predicted values obtained under the *FM* model and the *SP* model may be meaningful in this type of study. Consequently, an evaluation of the performance of the predictors derived under these three models for a wide range of conditions may be very helpful for practical applications. We consider such a comparison with the objective of selecting the predictor with smaller mean squared error (*MSE*).

The *ME*, *SP* and *FM* models can all be defined via a set of assumptions on the mean and on the covariance structure and do not require the specification of the form of the underlying distribution. Only the *FM* model links the finite population to the assumptions for the set of random variables that represent two-stage sampling (plus response error). When all variances are known and within-cluster variances are equal, Stanek and Singer (2004) show that the predictors of realized cluster latent values based on such a model have smaller *MSE* than those based on the other approaches. In practical situations, variances are rarely known and need to be estimated. We propose estimators for such variances and report simulation study results that compare the performance of empirical predictors of realized cluster latent values, providing guidance for the choice among the alternatives.

In Section 2 we present a brief review of the models and specify the corresponding predictors of sampled cluster latent values. We also propose empirical predictors based on variance components estimated from the sample. In Section 3 we describe technical details of the simulation study to compare the performance of these predictors for finite populations with different structures. Finally, in Sections 4 and 5 we present the simulation results and discussion, respectively. Programs and additional results are available at http://www.umass.edu/cluster/ed/.

[1] Although Stanek and Singer (2004) use the term “random permutation model”, we prefer “finite population mixed model” in order to avoid confusion with the SP random permutation model of Hedayat and Sinha (1991).