DOCUMENT RESUME

ED 385 567 TM 024 004

AUTHOR Longford, Nicholas T.

TITLE Comparison of Efficiency of Jackknife and Variance Component Estimators of Standard Errors. Program Statistics Research. Technical Report.

INSTITUTION Educational Testing Service, Princeton, N.J.

REPORT NO ETS-RR-92-24

PUB DATE Apr 92

NOTE 34p.

PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); Prediction; Research Design; Robustness (Statistics); *Sampling; *Scoring; Simulation; Statistical Bias; Surveys

IDENTIFIERS Bootstrap Methods; *Jackknifing Technique; Large Scale Programs; Resampling Techniques; *Variance (Statistical)

ABSTRACT

Large scale surveys usually employ a complex sampling design and as a consequence, no standard methods for estimation of the standard errors associated with the estimates of population means are available. Resampling methods, such as jackknife or bootstrap, are often used, with reference to their properties of robustness and reduction of bias. A method based on variance component models is proposed as an alternative to the jackknife procedure used for calculation of the standard errors for the subpopulation means of proficiency scores in a large scale survey of education in the United States. A simulation study provides evidence that the jackknife estimator for the standard error of the estimate of the mean is substantially less efficient than its variance component counterpart. The ultimate decision to use variance component methods should be based on the predicted (guessed) impact of the features of the data not accounted for by the variance component models. An appendix contains the scoring algorithm. Six tables present analysis results. (Contains seven references.) (Author/SLD)

RR-92-24

Comparison of Efficiency of Jackknife and Variance Component Estimators of Standard Errors

Nicholas T. Longford
Educational Testing Service

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 92-24

Educational Testing Service
Princeton, New Jersey 08541

Comparison of Efficiency of Jackknife and Variance Component Estimators of Standard Errors

Nicholas T. Longford
Educational Testing Service

Program Statistics Research
Technical Report No. 92-24

Research Report No. 92-24

Educational Testing Service
Princeton, New Jersey 08541

April 1992

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Abstract

Large scale surveys usua!ly employ a complex sampling design, and as a consequence no standard methods for estimation of the standard errors associated with the estimates of population means are available. Resampling methods, such as jackknife or bootstrap are often used, with reference to their properties of robustness and reduction of bias. We examine a method based on variance component models as an alternative to the jackknife procedure used for calculation of the standard errors for the subpopulation means of proficiency scores in a large scale survey of education in the U.S.A.

Keywords: Efficiency; Jackknife; Sampling design; Standard errors; Variance components.

1. Background and motivation

The National Assessment of Educational Progress (NAEP) is a large scale survey of U.S. primary and secondary schools. It employs a stratified three-stage clustered sampling design for students in various age/grade groups, and a complex partially balanced incomplete block design for the administered items. The item administration design enables collecting information about a large number of items without administering each item to every individual in the sample. The questionnaire items are divided into content areas (academic subjects) and, within subjects, into attitude and cognitive items. A common block of background items is administered to all the individuals.

For each content area an underlying proficiency (ability) scale is defined, and the scores on this scale are estimated from the responses to the cognitive items for all the students in the sample who have been administered at least one block of items from the content area. The proficiency scale is defined in such a way as to have, theoretically, the normal distribution with mean 250 and standard deviation 50. Each item has a limited number of response options, and for each cognitive item one response is correct. Results of the survey are published in the form of 'Summary Tables' which contain the sample (weighted) means of these proficiency scores, and the estimated standard errors for these means, for each combination of attitude item and response to it, cross-classified by the demographic background variables.

For example, the 1983-84 survey of 13-year-olds used a sample of approximately 31,000 students, each of whom was administered at least one of the 13 blocks of items pertaining to reading skills. For example, one of these 'reading' blocks (block N) was administered to 3,078 students. To the attitude item No. 4 of this block 2,139 students (approximately 70%) chose the response option A. One entry of the 'Summary Tables' contains an estimate of the mean proficiency of these students, and an estimate of the associated sampling variance. Most attitude items have five response options, and so the estimate of a typical entry in the Summary Tables is based on a small proportion of the total sample.

The sampling design involves 32 strata, within each of which a pair of primary sampling units (PSU's) is selected, with replacement. Schools are sampled within each selected PSU, and students are sampled within each selected school. The sampling procedures at each stage (PSU, school, student) are conditionally independent, given selection of the units at the higher level of aggregation. The (conditional) sampling probabilities are unequal, so as to oversample certain minority groups. The a priori (base) sampling weights were adjusted after the sampling procedure for non-response, and extremely large weights were trimmed so as to reduce the influence of the associated observations. Finally, the weights were adjusted by a process called poststratification to conform to certain population totals. We refer to these adjusted weights as poststratified weights.

Let Y_{hIJK} be the score of student K in school J within primary sampling unit (PSU) I of the stratum h. The population mean is defined as

$$\bar{Y} = \Sigma_{hIJK} Y_{hIJK} / \Sigma_{hIJ} N_{hIJ}, \qquad (1)$$

where N_{hIJ} is the number of students (of a particular age, or in a given grade) in the school (h, I, J). The mean for a subpopulation is defined similarly to (1), with N_{hIJ} replaced by the counts of students belonging to the subpopulation, within schools.

In NAEP the traditional ratio estimator for the (sub-) population means is used:

$$\bar{y} = \Sigma_{hijk}\, y_{hijk} w_{hijk} \,/\, \Sigma_{hijk}\, w_{hijk}, \tag{2}$$

where w_{hijk} are the poststratified weights, and the summations are over all the students in the sample and (if applicable) in the subpopulation. The sampling variance associated with this estimator is estimated by a jackknife method: For each stratum $h = 1, ..., H$ the h-pseudosample is created from the original sample by replacing the data for the first PSU in the stratum with the data from the other PSU in the stratum. The jackknife estimator of the sampling variance for the ratio estimator (2) is defined as the corrected sum of squares of the pseudosample means:

$$\hat{\sigma}^2 = \Sigma_h\, (\bar{y}_h - \bar{y})^2 \tag{3}$$

where $\bar{y}_h$ is the weighted mean for the h-pseudosample, with its weights adjusted for non-response in this pseudosample. The ratio estimator (2) itself is not jackknifed since it is believed to have satisfactory properties.

The jackknife procedures are computationally very extensive and cumbersome because they require calculation of the sampling weights adjusted for each jackknife pseudoanalysis associated with a stratum. In the case of NAEP the jackknife estimates of the standard errors for the population means, and for certain subpopulation means in particular, are known to have very poor sampling properties (Johnson, 1988).

Several researchers have proposed model-based estimation procedures for data from surveys that involve hierarchies (Malec & Sedransk, 1985, Aitkin & Longford 1986, and Battese, Harter & Fuller, 1988). The common feature of these variance component methods, considered from either a Bayesian or a likelihood prospective, is the modelling of the correlation structure of the observed data, or equivalently, the decomposition of the variation due to the levels of hierarchy induced by the sampling design.

The selected clusters at each level of the nesting hierarchy are a random sample from the respective populations of clusters (PSU's, schools, students), and so it is natural to represent the individual proficiency scores by the variance component model

$$y_{hijk} = a_h + b_{hi} + c_{hij} + e_{hijk}, \tag{4}$$

where the random terms b, c, e, form mutually independent samples from the normal distributions with means 0 and variances σ_3^2, σ_2^2 and σ_1^2, respectively. For the stratum means we consider two complementing assumptions:

A. They are unknown constants (fixed between-stratum differences).

B. They form a random sample from $N(\mu, \sigma_4^2)$ (random between-stratum differences).

The strata are set prior to sampling, and so the assumption A. is more appropriate. The assumption B. is attractive in that the 32 parameters a_h are replaced by just two, μ and σ_4^2. The original definition of the strata contains elements of arbitrariness, and that gives some credence to the assumption B.

Jackknife is a very general method, and it involves essentially no parametric assumptions. On the other hand, variance component methods are likely to be superior when the associated assumptions are satisfied, but in general they are much less robust than the jackknife procedures. The purpose of our study is to explore how and to what extent the jackknife procedures used in NAEP could be replaced by computationally more efficient methods, based on variance component analysis, that do not involve resampling.

The paper is organized as follows: In Section 2 we describe the datasets used for the study and compare the results of the jackknife and variance component analyses. In Section 3 the performance of the jackknife analysis is compared with the variance component analysis by means of two simulation studies. Artificial data were generated according to the variance component model (4), in order to evaluate the extent of the largest possible loss of efficiency in the jackknife procedures. Technical details are given in Section 4 and in the Appendix.

2. Data, procedures, and summary of results

From the 1983-84 assessment of reading data for the 13-year-olds who had been administered the block N of reading items were extracted. The jackknife procedure used in the operation of NAEP was replicated and the variance component models (4), with the assumptions A. and B., fitted for the data corresponding to the students with a specific response to a selected attitude item. We discuss the results for a representative set of item-by-response combinations, given in Table 1, that were selected in such a manner as to cover the entire range of the proportions of students that occur in the Summary Tables (approximately 10% - 100%).

TABLE 1 HERE

For estimation of the parameters in the variance component model (4) a modification of the Fisher-scoring algorithm of Longford (1987), adapted for unequal sampling weights, was used. The jackknife and variance component estimates of the standard errors for the estimates of the means are given in Table 2.

TABLE 2 HERE

The standard errors for the means using the variance component model with the assumption of fixed stratum-differences (A.) are very close to the jackknife standard errors. The largest discrepancy occurs for the case of 'all students', where the variance component standard error is about 10% higher. Assuming random stratum-differences leads to substantial overestimates of the standard errors, almost 30% in the case of 'all students'. The two variance component models result in identical estimates of the standard errors for the cases 6B (response B to the item 6) and 9C for which the estimate of the stratum-level variance σ_4^2 is equal to 0.

For variance component analysis we use the parametrization

$$(\sigma_1, \tau_2, \tau_3, \tau_4),$$

where $\tau_h = \sqrt{\sigma_h^2/\sigma_1^2}$ is the square root of the ratio of the h-level and the student-level variances. Thus, the variance of an observation is equal to $\sigma_1^2(1 + \tau_2^2 + \tau_3^2 + \tau_4^2)$. The estimates of the variance components are given in Table 3. These results indicate that the between-school variation (within-PSU, or level 2) accounts for between 10% (all students) and 20% (9C) of the total variation. The between-PSU (level 3) variance is substantially smaller.

TABLE 3 HERE

The estimates of the variances for the model assumptions A. and B. are identical except that the estimates for the between-stratum variation in model B., are replaced by the 31 stratum-contrasts for model A.

The main implication of these results is that the standard errors of the (sub-)population means obtained from variance component model fits can be used instead of the computationally more intensive jackknife procedures. The main advantage of the variance component approach is that no resampling weights need to be calculated, and so the process of reporting of the results could be considerably streamlined.

3. Simulations

In this Section we discuss the issue of efficiency of the jackknife estimation method. For the model assumptions (4) the direct maximum likelihood method is asymptotically fully efficient, and so it is reasonable to assume that the relative efficiency of the jackknife and variance component methods for model (4) provides the most unfavorable comparison for the jackknife.

3.1 Jackknife vs. variance components

Data sets were generated according to the model (4) with the assumption of fixed stratum-differences. Since all the estimators of the variance components are translation invariant, our results are unaffected by the actual choice of the stratum-differences, and therefore they can be set identically to zero. In order to simplify the study further, we generated the following stratum/PSU/school design: For a given data set design (such as 4A, see Table 1) we generate a 'simulation' design by rounding the within-school totals of weights — these integers then represent the numbers of students within the schools in the simulation design. The design at the higher levels, i.e., the clustering of schools within PSU's and the pairs of PSU's within strata, is left intact. Equal 'simulation' weights are assigned to each observation.

We report only the results for the imputed values of the variance components 1. (students), 0.12 (schools), and 0.03 (PSU's); they are representative of the results for other realistic values of the variance components. The model (4) and all the estimators used are invariant with respect to linear transformations, and therefore the student-variance can be set to an arbitrary positive value and the population mean to any real value. The other two variances are close to the values of the estimates in the real data. Two hundred replicates of the simulation datasets based on the data for all the students and for the combination 4A were generated. In order to informally confirm the generalizability of the results the variance components for schools in the range 0.04 - 0.20 and for PSU's in the range 0.01 - 0.10 were also used.

The results of the simulations indicate that the jackknife estimator of the mean does not provide any improvement over the arithmetic average, but the variance component estimator is appreciably more efficient. The variance component estimator for the standard error of the mean is far superior to its jackknife counterpart. The relevant results of the simulation study are summarized in Table 4.

TABLE 4 HERE

The Table contains three pairs of rows, corresponding to the arithmetic mean (i.e., assuming simple random sampling), the variance component method and the jackknife method. Within each pair the first row corresponds to the estimator of the mean and the second to the associated estimator of the standard error. The row 'VC.GM.' represents the standard error of the arithmetic mean under the assumption of the variance component model. The bias of each estimator of the standard error for the mean can be assessed by comparing the mean of the model-based estimates for the standard error with the sampling standard deviation of the corresponding estimates for the mean. For example, for the 'all students' dataset the sampling standard deviation of the ordinary mean (0.0371) is 1.87 times higher than the average square root of the mean square errors (.0198). The corresponding ratio for the dataset 4A is 1.67.

The jackknife estimates of the mean are closer to the ordinary means than the variance component estimates. The jackknife estimator of the standard error for the mean has very little bias (compare mean of JK.SE. with the sampling standard deviation of JK.M.), and it also estimates the sampling standard deviation of the ordinary mean (G.M.) without any observable bias. The variance component estimator for the sampling standard deviation of the ordinary mean (VC.GM.) overestimates the sampling standard deviation of the ordinary mean by about 4%.

The variance component estimator for the mean (V.C.M.) is appreciably more efficient than the jackknife estimator. Its sampling standard deviation is lower than the sampling standard deviation for the jackknife or the ordinary mean by 9% (all students) and 7% (4A). Note however, that the estimate of the sampling standard deviation for the variance component mean is biased (compare the mean of VC.SE with the sampling standard deviation for the V.C.M.), it has a positive bias of about 6% for both data sets.

The sampling standard deviations for the variance component estimators are substantially smaller than their jackknife counterparts. The sampling standard deviations for the VC.SE. and VC.GM. are about 33% (all students) and 20% (4A) smaller than the corresponding values for JK.SE.

We conclude that, contingent on appropriateness of the variance component model (4), estimation of the mean could be moderately improved (i.e., lower mean squared error) by application of variance component methods, but substantial improvement in the sampling properties of the estimates of the associated standard errors would result. The additional benefit would be that the resampling weights could be dispensed with.

3.2 Lumpy data

The Summary Tables contain a large number of entries related to subpopulations, such as minorities, which constitute only a small proportion of the target population, and they may be very unevenly distributed across the strata. The standard errors obtained by the jackknife procedures are subject to substantial sampling variation, and their estimation is probably very inefficient (Johnson, 1988). On the other hand the asymptotic properties of the maximum likelihood estimators using the variance component models may not hold for small datasets.

In order to compare the jackknife and variance component methods for such 'lumpy' data we have generated several data sets from the artificial dataset of 'all students' by the following two-stage sampling design:

1. For each stratum generate a value p_1 from $U(0.2, 0.5)$ (uniform distribution on the interval 0.2 - 0.5). Then, for the schools in the stratum, include the whole school in the dataset with (stratum-specific) probability p_1.

2. For the included schools: For each school generate a value p_2 from $U(.1, .5)$, and include a student from the school in the dataset with (school-specific) probability p_2.

We discuss the results for three such datasets, contaiaing 368 students in 98 schools, 329 students in 109 schools and 284 students in 85 schools, respectively. In these datasets between 8-12 PSU's and 1-3 strata were not represented at all. For illustration, the nesting design for one of these datasets is given in Table 5; most PSI "s are represented by fewer than 10 students, although 5 schools have 10 or more students in the dataset.

TABLE 5 HERE

The proficiency scores were generated by the variance component model (4) with the variances $\sigma_1^2 = 1.$, $\sigma_2^2 = 0.12$, $\sigma_3^2 = 0.03$ and $\sigma_4^2 = 0$, and mean $\mu = 0$. Results of the simulation stndy using 200 replicates are given in Table 6. Table 6 has the same format as Table 4, but in addition it contains summary statistics for the jackknife and variance component estimators of sampling variance (see (5) below). We see that the mean squared error (M.S.E.) is biased and the jackknife estimatc of the standard error agrees with the sampling standard deviation of both the jackknife and the ordinary mean estimators. The variance component estimator of the mean is only marginally more efficient.

The variance component estimator of the standard error for the mean is more biased than the jackknife estimate, but its sampling standard deviation is about twice as small as that of the jackknife.

TABLE 6 HERE

The number of degrees of freedom associated with the estimators of the standard error can be estimated by the formula

$$\frac{2(\text{estimate of the variance})^2}{\text{sampling variance of the squared standard error}} \qquad (5)$$

derived by matching the moments of the χ^2 distribution. These value are given in Table 6. The variance component estimator of the standard error has 35-40 more degrees of freedom than its jackknife counterpart. Note that in the jackknife 31 degrees of freedom is the upper limit for *any* data set.

We conclude that in small and lumpy datasets the variance component estimator of the standard error for the mean is likely to be much more efficient than the jackknife estimator probably even in presence of features that to a moderate extent violate the assumptions of the variance component model.

4. Computational details

In the jackknife analysis we define a pseudosample corresponding to each stratum, and carry out the 'basic' analysis for each sample. In our case the basic analysis consists of calculation of the weighted mean (2), and the pseudosample for stratum h is generated from the original data set by deletion of the first PSU of the stratum and replacing it with the second PSU of the stratum, with unaltered sampling weights. The sampling weights are then adjusted for poststratification. Let the weighted mean of the pseudosample h be $\overline{y}_h$, and the weighted mean for the original data set $\overline{y}$. The jackknife estimator for the mean (1) is given by

$$\overline{y}^* = \Sigma_h \{H\overline{y}_h - (H - 1)\overline{y}\}$$

(H is the number of strata, 32), and the sampling variance of this estimator is estimated by (3). For further details we refer the reader to Beaton et al. (1988), Ch. 14.2.

The Fisher scoring algorithm for variance component analysis requires formulae for the Jacobian and the expectation of the Hessian associated with the estimated parameters. For easier description we consider first the case of equal weights. The log-likelihood for a set of observations with equal sampling weights is given, apart from an additive constant, by the formula

$$-2\log \lambda = \log \det(\mathbf{V}) + \mathbf{e}^T\mathbf{V}^{-1}\mathbf{e},$$

where $\mathbf{V}$ is the variance matrix for the observations and $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$ is the vector of residuals. The determinant and the inverse of the variance matrix can be evaluated efficiently, and without numerical inversion of any matrices by the recursive algorithm described in Longford (1987) where the formulae for the Jacobian and Hessian are derived. Details are given in Appendix. The computational procedure is based on the counts of students within schools, the within-school totals of proficiencies and the sample total of squares of proficiencies. The weighted version of the algorithm uses the same formulae, with the counts of students replaced by totals of weights, the totals of proficiencies by the corresponding weighted totals and the sum of squares of proficiencies by the weighted sum of squares. The sampling weights are normalized (multiplied by a constant) so that the sample total of the normalized weights is equal to the number of students in the sample.

The adopted parametrization has the advantage that the estimate of the elementary variance σ_1^2 is obtained at each iteration by setting the Jacobian to zero:

$$2\,\partial(\log \lambda)/\partial\sigma_1^2 = -n/\sigma_1^2 + e^T W^{-1} e/\sigma_1^4 = 0,$$

where $\mathbf{W} = \sigma_1^{-2}\mathbf{V}$. Note that in the τ-parametrization $\mathbf{W}$ does not depend on σ_1^2. Instead of the variance ratios τ_2^2, τ_3^2 and τ_4^2 their respective square roots, τ_2, τ_3 and τ_4, are estimated; The Jacobian and Hessian are adjusted by the chain rule. The main advantage of estimating (ratios of) standard deviations instead of (ratios of) variances is that negative estimates of the variances are avoided. Also, the standard errors obtained from the inverse of the estimated expected information matrix are easier to interpret because negative standard deviations in a confidence interval correspond to positive variances.

In the model (4) the constant μ can be replaced by a linear predictor, such as one allowing different within-stratum means. In general, addition of an explanatory variable will reduce the variance components (or leave them unchanged). In the model with random strata a variable defined for strata will leave the PSU-, school- and student-level variances unchanged, and can reduce only stratum-level variance. The stratum factor (categorical variable with 32 categories) will saturate the stratum-level variance, and will result in a zero stratum-level variance component. Thus the overall mean in the model (4) with fixed stratum-differences can be estimated by applying the model (4) with random stratum-differences, and then setting the stratum variance to zero. Direct estimation of the 32 stratum-means by the Fisher scoring method would involve iterative inversion of 32 x 32 matrices, a substantial burden compared to the estimation of the variance components. As an alternative the ordinary within-stratum means could be imputed for them in variance component estimation.

5. Summary

The reported study has demonstrated that the computationally extensive procedures based on the jackknife method can be replaced by variance component methods which do not involve resampling. The differences between the jackknife and variance component-based estimates of the (sub-) population means are of no practical importance since they are comparable to rounding errors. For small data sets (with 1000 or fewer subjects) the jackknife and the variance component-based estimates of the standard errors for the corresponding estimates of population means are almost identical, but for subpopulation means the jackknife standard errors are substantially less efficient than their variance component counterparts. For larger data sets some differences arise, but they cause no noticeable changes in the Summary Tables. The efficiency of the jackknife and variance component methods can be compared only on simulated data sets, such as those generated by a variance component model. The simulation study described in Section 3 provides evidence that the jackknife estimator for the standard error of the estimate of the mean is substantially less efficient than its variance component counterpart. Small proportion of this loss can be attributed to the difference in efficiency of the jackknife and variance component estimators of the mean. The ultimate decision to use variance component methods should be based on the predicted (guessed) impact of the features of the data not accounted for by the variance component models, such as the nature of the poststratified sampling weights, and possible variation of the variance components across the strata. The impact of these features in the analyses of the studied data sets appears to be ignorable, but only a study extended to the entire variety of dataset designs involved in NAEP could arbitrate whether these features can be ignored throughout, and the jackknife resampling weights made obsolete. The gain in efficiency by using variance component analysis is most striking in small 'lumpy' datasets because the jackknife estimator of the standard error ignores the within-PSU information.

The additional information provided by the variance component analysis consists of the estimates of the variance components. The school-level variance is of primary interest; it provides a description of school heterogeneity. Future alterations in the design of the entire survey could be easier to plan, with optimality of inference as a goal, and some results, such as the standard errors for population means, could be predicted prior to data collection using past (or imputed) values of the variance components.

APPENDIX

Fisher scoring algorithm for variance component estimation

Suppose the observations for the subjects (students) are in lexicographic order, so that their variance matrix $\mathbf{V}$ is block-diagonal and equal to

$$\mathbf{V} = \sigma_1^2\mathbf{W} = \sigma_1^2(\mathbf{I} + \tau_2^2\mathbf{J}^{(2)} + \tau_3^2\mathbf{J}^{(3)} + \tau_4^2\mathbf{J}^{(4)})\,, \qquad \text{(A.1)}$$

where all the matrices have sizes n x n, $\mathbf{I}$ is the unit matrix and $\mathbf{J}^{(p)}$, $p = 2,3,4$, are the respective incidence matrices for schools, PSU's and strata; the element (r_1, r_2) of $\mathbf{J}^{(p)}$ is equal to 1 if the students r_1 and r_2 belong to the same unit at the level p, and is equal to 0 otherwise. Fixed stratum-differences correspond to $\tau_4^2=0$. Let $\mathbf{1}_{hij}^{(2)}$, $\mathbf{1}_{hi}^{(3)}$, $\mathbf{1}_{h}^{(4)}$ and $\mathbf{1}$ denote the respective n x 1 indicator vectors for a school, a PSU, a stratum, and for the whole sample, so that

$$\mathbf{J}^{(2)} = \Sigma_{hij}\,\mathbf{1}_{hij}^{(2)}\,\mathbf{1}_{hij}^{(2)T}, \quad \mathbf{J}^{(3)} = \Sigma_{hi}\,\mathbf{1}_{hi}^{(3)}\,\mathbf{1}_{hi}^{(3)T}, \quad \mathbf{J}^{(4)} = \Sigma_{h}\,\mathbf{1}_{h}^{(4)}\,\mathbf{1}_{h}^{(4)T},$$

$$\mathbf{1}_{hi}^{(3)} = \Sigma_j\,\mathbf{1}_{hij}^{(2)}, \quad \mathbf{1}_{h}^{(4)} = \Sigma_i\,\mathbf{1}_{hi}^{(3)}, \quad \text{and} \quad \mathbf{1} = \Sigma_h\,\mathbf{1}_{h}^{(4)}. \qquad \text{(A.2)}$$

In order to simplify and streamline the notation we will use the symbol $\mathbf{\Sigma}_{(p)}$ for the summation over all units at the level p, and the dot notation; for example,

$$\mathbf{J}^{(p)} = \sum_{(p)} \mathbf{1}_{\bullet}^{(p)}\mathbf{1}_{\bullet}^{(p)T}.$$

The log-likelihood for the model (4) is given by the formula

$$-2\log\lambda = n\log\sigma_1^2 + \log\det(\mathbf{W}) + \mathbf{e}^{T}\mathbf{W}^{-1}\mathbf{e}/\sigma_1^2 \qquad \text{(A.3)}$$

where $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$ is the vector of residuals, i.e., the differences of the observed and model values. The vector $\boldsymbol{\mu}$ may vary across the observations, e.g., related to a set of explanatory variables by a linear formula, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is a design matrix of known constants, and $\boldsymbol{\beta}$ a vector of (known or unknown) location parameters. In most cases we consider the case of constant predictor $\boldsymbol{\mu}$ ($= \mu\,\mathbf{1}$) or stratum-specific means μ_h.

The estimate of the elementary-level variance is obtained by setting the derivative of (A.3) with respect to σ_1^2 to zero:

$$n/\sigma_1^2 - \mathbf{e}^T\mathbf{W}^{-1}\mathbf{e}/\sigma_1^4 = 0,$$

which has, for a given $\mathbf{W}$, the unique solution $\sigma_1^2 = \mathbf{e}^T\mathbf{W}^{-1}\mathbf{e}/n$.

The first partial derivatives with respect to τ_p^2, $p=2,3,4$, are equal to

$$\partial(\log \lambda)/\partial\tau_p^2 = -\tfrac{1}{2} \sum_{(p)} \{\mathbf{1}_{\cdot}^{(p)T}\mathbf{W}^{-1}\mathbf{1}_{\cdot}^{(p)} - (\mathbf{e}^T\mathbf{W}^{-1}\mathbf{1}_{\cdot}^{(p)})^2/\sigma_1^2\}, \qquad \text{(A.4a)}$$

and the expectations of the second derivatives ($q \geq p$) are

$$E\{\partial^2\log \lambda/\partial\tau_p^2\partial\tau_q^2\} = -\tfrac{1}{2}\, \mathrm{tr}\{\mathbf{J}^{(p)}\mathbf{W}^{-1}\mathbf{J}^{(q)}\mathbf{W}^{-1}\} = -\tfrac{1}{2} \sum_{(q)} \sum_{(p|q)} (\mathbf{1}_f^{(q)T}\mathbf{W}^{-1}\mathbf{1}_g^{(p)})^2, \qquad \text{(A.4b)}$$

where the double summation is over all units f at the level q and all its subunits g at the level p.

The maximum likelihood estimator for the regression parameters is given by the generalized least squares formula

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\, \mathbf{X}^T\mathbf{W}^{-1}\mathbf{y},$$

where $\mathbf{X}$ is the (regression) design matrix, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. For the case of no explanatory variables we have the estimate of the grand mean

$$\hat{\mu} = (\mathbf{1}^T\mathbf{W}^{-1}\mathbf{1})^{-1}\, \mathbf{1}^T\mathbf{W}^{-1}\mathbf{y} \qquad \text{(A.5)}$$

with the information $\sigma_1^2\, \mathbf{1}^T\mathbf{W}^{-1}\mathbf{1}$.

Since **W** is a matrix of large size it is important to have efficient algorithms for computation of expressions involving $\mathbf{W}^{-1}$. We define $\mathbf{W}_1 = \mathbf{I}$ and

$$\mathbf{W}_p = \mathbf{W}_{p-1} + \tau_p^2 \mathbf{J}^{(p)},$$

$p = 2,3,4$, so that $\mathbf{W}_4 = \mathbf{W}$. For the inverses of these matrices we have the recursive formula

$$\mathbf{W}_p^{-1} = \mathbf{W}_{p-1}^{-1} - \mathbf{W}_{p-1}^{-1} \sum_{(p)} \{\mathbf{1}_\cdot^{(p)}\, \mathbf{1}_\cdot^{(F)T}\, \tau_p^2/(1 + \tau_p^2\, \mathbf{1}_\cdot^{(p)}\mathbf{W}_{p-1}^{-1}\mathbf{1}_\cdot^{(p)})\}\, \mathbf{W}_{p-1}^{-1}. \qquad \text{(A.6)}$$

We define

$$C_\cdot^{(p)} = \mathbf{1}_\cdot^{(p)T}\mathbf{W}_{p-1}^{-1}\mathbf{1}_\cdot^{(p)}$$

and

$$E_\cdot^{(p)} = \mathbf{e}^T\mathbf{W}_{p-1}^{-1}\mathbf{1}_\cdot^{(p)}, \qquad \text{(A.7)}$$

where the dot . stands for a unit at the level p, and

$$C = \mathbf{1}^T\mathbf{W}^{-1}\mathbf{1},\ E = \mathbf{e}^T\mathbf{W}^{-1}\mathbf{1}.$$

We have $C_\cdot^{(p)} = n_\cdot^{(2)}$ (number of students from school • in the sample), and $E_{hij}^{(2)} = \Sigma_k\, e_{hijk}$ (sum of the residuals within school hij). The inversion formula (A.6) implies the identities

$$D_{hi}^{(3)} = \Sigma_j\, D_{hij}^{(2)}/(1 + \tau_2^2 C_{hij}^{(2)}),$$

$$D_h^{(4)} = \Sigma_i\, D_{hi}^{(3)}/(1 + \tau_3^2 C_{hi}^{(3)}),$$

and

$$D = \Sigma_h D_h^{(4)}/(1 + \tau_4^2 C_h^{(4)}), \tag{A.8}$$

where D stands for either C or E.

It is easy to show that

$$\log\det(\mathbf{W}) = \sum_{p=2}^{4} \sum_{(p)} \log(1 + \tau_p^2 C_{\bullet}^{(p)}) \tag{A.9}$$

and

$$\mathbf{e}^T\mathbf{W}^{-1}\mathbf{e} = \sum_{p=2}^{4} \sum_{(p)} \{E_{\bullet}^{(p)}\}^2 \tau_p^2/(1 + \tau_p^2\, C_{\bullet}^{(p)}). \tag{A.10}$$

These formulae enable efficient calculation of the log-likelihood (A.3). All the quadratic forms required for (A.4a), (A.4b) and (A 5) can be calculated directly from the constants $C_{\bullet}^{(p)}$ and $E_{\bullet}^{(p)}$ using (A.6). The partial derivatives with respect to the square roots of the variances are calculated from (A.4a) and (A.4b) using the chain rule.

The Fisher scoring algorithm is an iterative procedure, and as such it requires initial values for all the estimated parameters. For the regression parameters (the population mean) the ordinary least squares (arithmetic mean) provides a suitable initial solution, and for the variance components any non-iterative procedure which provides positive values is suitable. We have used a naive moment estimate which in the models fitted turned out to have values between 50-200% of the maximum likelihood estimate. The Fisher scoring algorithm required between 6-12 iterations. The iterations were terminated when both the change of the -2 log-likelihood and of each τ_p were smaller than 10^{-4}.

The estimator of the sampling variance of the arithmetic mean under the variance component model (VC.GM. in Table 4) is equal to $\mathbf{1}^T\mathbf{V}\mathbf{1}/n^2$; its evaluation is straightforward using (A.1).

Adaptation for sampling weights.

Formally, the variance matrix **V** in (A.1) can be replaced by

$$\mathbf{V} = \sigma_1^2 \mathbf{H}^{½}\mathbf{W}\mathbf{H}^{½},$$

where **H** is a diagonal matrix of sampling weights. All the formulae (A.4) - (A.8) carry over directly after redefining the within-school scalars in (A.7) as

$$C_{hij}^{(2)} = \Sigma_k H_{hijk}$$

and

$$E_{hij}^{(2)} = \Sigma_k e_{hijk} H_{hijk} \,. \qquad \text{(A.11)}$$

An iteration of the algorithm starts with the scalars (A.11) from which the level-3 and level-4 totals $C^{(3),}$ $E^{(3)}$, $C^{(4)}$, a.d $E^{(4)}$, as well as sample totals C and E are calculated using (A.8). From these scalars the items for the Jacobian and Hessian of the Fisher scoring algorithm are calculated using (A.6).

REFERENCES

Aitkin, M.A. & Longford, N.T. (1986). Statistical modelling issues in school effectiveness studies. *J. R. Statist. Soc.* A, **149**, 1-43.

Battese, G.E., Harter, R.M. & Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Am. Statist. Ass.*, **83**, 23-36.

Beaton, A.E. et al. (1988). Expanding the New Design. The NAEP 1985-86 Technical Report. Educational Testing Service, Princeton, NJ.

Johnson, E. (1988). Considerations and techniques for the analysis of NAEP data. ETS Research Report RR-88-49, Educational Testing Service, Princeton, NJ.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, **74**, 817-27.

Malec, D. & Sedransk, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *J. Am. Statist. Ass.*, **80**, 897-840.

Scott, A. & Smith, T.M.F. (1969). Estimation in multi-stage surveys. *J. Am. Statist. Ass.*, **64**, 830-840.

Table 1. The data sets used in the study, with numbers of students and schools.

Item	Response	Students	Schools
6	B	309	225
9	C	560	299
5	E	1,038	325
8	A	1,312	341
4	A	2,240	379
All Students		3,076	392

Table 2. Comparison of the jackknife and variance component estimates.

The means from variance component analysis are quoted only for model A.; for model B. they differ by less than .0005.

Item-Response		JACKKNIFE		VARIANCE COMPONENT		
					St. Error	
		Mean	St. Error	Mean	A.	B.
6	B	250.76	1.694	250.98	1.685	1.685
9	C	257.19	1.768	257.22	1.730	1.730
5	E	264.82	1.265	264.39	1.272	1.664
8	A	252.42	1.252	252.08	1.241	1.388
4	A	254.81	1.263	255.12	1.299	1.630
All Students		253.38	1.046	253.51	1.154	1.350

Key: A. - fixed stratum differences

B. - random stratum differences

Table 3. Estimates of the variance components.

The first line in every cell contains the estimate of the variance, the second line the square root of this estimate, and the third line (in parentheses) the standard error for the square root.

Case	σ_1^2 (students)	τ_2^2 (schools)	τ_3^2 (PSU's)	τ_4^2 (strata)
6B	9.83 3.136	.1428 .3379 (.0736)	.0000 .0000 (.1921)	.0000 .0000 (.1919)
9C	10.28 3.206	.2203 .4693 (.0757)	.0083 .0909 (.2074)	.0000 .0000 (.1925)
5E	10.45 3.237	.1575 .3968 (.0512)	.0000 .0000 (.2327)	.0287 .1693 (.0634)
8A	10.39 3.223	.1414 .3761 (.0483)	.0076 .0870 (.1480)	.0099 .0995 (.0959)
4A	10.94 3.308	.1457 .3817 (.0483)	.0332 .1823 (.1480)	.0286 .1692 (.0959)
All students	11.55 3.398	.1223 .3497 (.0319)	.0234 .1530 (.0628)	.0136 .1168 (.0672)

Table 4. Results of the simulation study for the 'all students' and the 4A design. Sampling means and standard deviations for various estimators of the mean and of the standard error (200 replicates).

Estimator	All students		4A	
	Mean Estimate	Sampling St. Dev.	Mean Estimate	Sampling St. Dev.
G.M.	-.00448	.03706	-.00670	.0386
M.S.E.	.01979	.00029	.02321	.00038
V.C.M.	-.00597	.03392	-.00625	.03631
VC.SE.	.03599	.00364	.03848	.00428
VC.GM.	.03851	.00375	.04059	.00446
JK.M.	-.00436	.03725	-.00651	.03895
JK.SE.	.03762	.00543	.03818	.00544

Key:

G.M. ... ordinary (arithmetic) mean

M.S.E. ... square root of the mean squared deviation from G.M. (the simple random sampling estimator of the standard error)

V.C.M. ... variance component (ML) estimator of the mean

VC.SE. ... estimator of the asymptotic standard error of V.C.M.

VC.GM. ... the estimator of the sampling standard deviation of G.M. given the variance component model (4), A.

JK.M. ... the jackknife estimator of the mean

JK.SE. ... the jackknife estimator of the standard error for the mean

Table 5. Simulated sampling design of a 'lumpy' dataset.

Counts of students within schools (368 students from 98 schools). For example, the second PSU of the stratum 2 has 3 schools in the dataset, with 6 students in one school and one each in the other two schools.

Stratum	First PSU				Second PSU			
1	5				1			
2	1				6	1	1	
3	6	4			9	2		
4	10	1	2	1	6			
5	1	1	1		2	2	3	1
6	5				3	4		
7	5	4			17			
8	4	8	17		7	2	3	
9	2	2						
10	7	3			1	2		
11	12	9	2		3	5		
12	1	2			2			
13	5				4	6		
14	7				1			
15	4	9	4		1			
16	5							
17	2	4			4			
18	4	1			9			
19	3	2	2		4	2		
20	2				1	1	1	
21	1				2			
22					1	6		
23	3	9			3			
24	10							
25	1	7	1		1	8		
26					1	1		
27	3	3			3	2		
28					2			
29	1	5	1					
30					1	5		

Table 6. Results of the simulation study for 'lumpy' data.

Sampling means and standard deviations for various estimators of the mean and of the standard errors (200 replicates).

	368 Students 98 Schools		329 Students 109 Schools		284 Students 85 Schools	
Estimator	Mean Estimate	Sampling St. Dev.	Mean Estimate	Sampling St. Dev.	Mean Estimate	Sampling St. Dev.
G.M.	.00273	.07478	.00966	.07975	-.00007	.07892
M.S.E.	.05601	.00215	.05913	.00237	.06358	.00268
V.C.M.	-.00078	.07110	.00838	.07467	-.00048	.07599
VC.SE.	.07301	.00649	.07480	.00754	.08159	.00768
VC.S2.	.00537	.00096	.00565	.00114	.00672	.00127
VC.GM.	.07740	.00885	.08058	.01077	.08636	.01034
JK.M.	.00268	.07498	.00993	.07923	.00022	.07930
JK.SE.	.07603	.01396	.07910	.01751	.08286	.01283
JK.S2	.00598	.00223	.00656	.00315	.00703	.00220
JK.DF.	14.4		8.7		20.4	
VC.DF.	62.2		48.7		56.3	

Key: See Table 4, and:

VC.S2 . . . estimator of the asymptotic variance of V.C.M.

JK.S2. . . . the jackknife estimator of the variance of JK.M.

JK.DF. . . . the estimated number of degrees of freedom of the jackknife estimator of the sampling variance

VC.DF . . . the estimated number of degrees of freedom of the variance component estimator of the sampling variance