Statistica Sinica **8**(1998), 1175-1188

A CONSTRAINED MINQU ESTIMATOR OF CORRELATED RESPONSE VARIANCE FROM UNBALANCED DATA IN COMPLEX SURVEYS

Sujuan Gao and T. M. F. Smith

Indiana University School of Medicine and Southampton University

Abstract: This paper is concerned with the estimation of correlated response variance from unbalanced data in complex surveys. The Minimum Variance Quadratic Unbiased Estimator (MINQUE) proposed by Rao (1971, 1972) is often used for variance components estimations. However, for unbalanced data the equations for obtaining the MINQU estimates are very difficult to solve. In this paper we propose a constrained MINQU estimator by substituting robust unbiased estimates for the random residual errors in the MINQUE equations to obtain estimates of the correlated response variance. We demonstrate through numerical and empirical studies that the constrained MINQUE is efficient compared to the full MINQU estimator. We also point out that the constrained MINQU estimator can be used in other cases where reducing the complexity of the MINQUE equations is desired.

Key words and phrases: Correlated response variance, Minimum Variance Quadratic Unbiased Estimator (MINQUE).

1. Introduction

In most sample surveys the responses are subject to measurement errors which depend on the nature of the survey and on the methods employed for collection, recording and processing, where the term measurement error is used as a generic term to cover all errors other than sampling errors. It is now recognized that these errors may be a major component of total survey error, and that effective control of the survey process requires that wherever possible all sources of error should be identified and their magnitude estimated. Systematic errors which cause biases can only be estimated if there is information on the "true" values, and these are rarely available. In this paper we are concerned with the estimation of the random errors that arise during the survey process.

The fundamental work on the estimation and control of measurement errors was carried out at the Bureau of the Census by Morris Hansen and his colleagues (see for example, Hansen, Hurwitz and Bershad (1961)). The Bureau of the Census model partitions random measurement error into correlated errors, arising from the use of common operators, such as interviewers, supervisors and coders, and uncorrelated, or simple, errors arising from the inevitable random errors

that may occur in any complex process such as a survey or a census. Operators common to several sampling units impose a clustering structure on the units which leads to an additional contribution to the total variance of a sample mean in the form $\sigma^2(1+(\bar{m}-1)\rho_0)$ where σ^2 is the random error associated with each sampling unit, $\bar{m}$ is the average number of units processed by each operator and ρ_0 is the intra-cluster correlation induced by the operator. If $\bar{m}$ is large then even small values of ρ_0 can lead to a substantial increase in the total variance. This has been the motivation for first measuring, and then controlling, the various sources of measurement error. For a recent review of measurement error effects on the analysis of survey data see Biemer and Trewin (1997) in Section E of Survey Measurement and Process Quality edited by Lyberg et al. (1997).

Estimation of measurement error variances requires replication. For simple errors it is usually impossible to replicate procedures independently and these errors are therefore completely confounded with the values of the observations, since only the sum of the observation and the measurement error can be observed. For correlated errors the confounding structure can be broken if it can be assumed that operator effects are fixed conditional on the operator, and two or more operators are employed within homogeneous groups of sampling units. A common survey design is to select two primary sampling units (psu) from within strata and to allocate one interviewer to each psu. Interviewers are then confounded with psus. Employing two interviewers within each psu breaks this confounding, and an ANOVA can be carried out within each stratum and the results can then be pooled across interpenetrated strata.

The cost of implementing an interpenetrated design is believed to be high, although if implemented routinely this need not be so. High cost means that only a small sample of strata may be used for the interpenetration experiment, with a consequent inefficiency in the estimation of the operator component of variance. In the balanced case, with equal workloads per operator, Fellegi (1974) demonstrated how the information on the sum of sampling and measurement errors from the noninterpenetrated groups could be combined with the information on the sampling variance from the interpenetrated groups to give a second estimator of the operator variance. He then argued that an average of the two estimators would often be better than the ANOVA estimator alone. The validity of his analysis depends on being able to select the interpenetrated groups (pairs of psus) at random from the set of all such sample groups. This random selection converts the stratum fixed effect into a random effect in a marginal analysis which averages over all possible selections.

Biemer and Stokes (1985) generalized Fellegi's approach to cover unequal workloads within groups and unequal variances for each group. Using a linear model approach as in Hartley and Rao (1978), they employed the synthesis version of MINQUE, MINQUE(0), due to Hartley, Rao and LaMotte (1978) as the

first estimator in the unbalanced case, and showed that this depended on the interpenetrated groups only. Essentially this is an alternative unbiased estimator to an ANOVA estimator. Their second estimator is a simple extension of Fellegi's second estimator. They also proved that under the assumption of normality the two estimators are independent and that an optimally weighted estimator would have smaller variance than either estimator individually. However, the optimal weights would usually be unknown. They gave explicit expressions for the estimators for the balanced case of equal achieved workloads. Their analysis also requires that the interpenetrated units be chosen at random.

The fact that a general estimation procedure such as MINQUE does not utilise the information in the noninterpenetrated groups is explained by the fact that the stratum effects are treated as fixed effects. Kleffe, Prasad and Rao (1991) extended the Biemer and Stokes model to encompass the complete set of sample groups by assuming that the group means were random variables following a components of variance model of the form proposed by Scott and Smith (1969). This model provides a link between the interpenetrated and the noninterpenetrated groups analogous to that provided by randomisation. They estimated the operator variance using the general MINQUE(α) procedure due to Rao (1971, 1972). Under the assumption of normality their estimator will be the locally best invariant unbiased estimator if the prior values (α) are correctly specified. They showed how to compute the estimator in the general unbalanced case of unequal workloads and unequal variances within groups, but recognized that the MINQU estimating equations would be difficult to solve if there were a large number of groups. They derived an explicit estimator for the balanced case, with both equal workloads and equal variances, and compared the efficiency of their new estimator with earlier estimators.

This review of the work which has followed Fellegi's original proposal has highlighted two issues. The first is that of linking the information in both interpenetrated and noninterpenetrated groups when the sampling effects are fixed stratification effects. The two alternatives are either to select groups at random and to average over the distribution of all such selections, or to model the group means and to carry out the analysis conditional on the given selection. Since the object of the exercise is analytic rather than descriptive we favour the modelling approach. This gives much more freedom to the selection of groups and to the allocation of operators to groups provided always that the allocation is uninformative in the sense of not being related to the effect being estimated. Randomisation guarantees that the allocation will be uninformative. The second issue is the technical one of solving the estimating equations for a class of realistic models which should encompass the unbalanced case with unequal variances at the final stage and unequal workloads. It is the unequal variance assumption

that makes the computations difficult. Empirical evidence from Rao, Kaplan and Cochran (1981) shows that it is inefficient to assume equality of variances when in fact they are unequal. In this paper we adopt the general model of Kleffe et al. (1991) and propose a new form of constrained MINQU estimator which can handle the computations in the unbalanced case more easily.

The paper is organized as follows. In Section 2 we propose the constrained MINQU estimator using the model assumption of Kleffe et al. (1991). In Section 3 we demonstrate that the constrained MINQU estimator performs well relative to the full MINQU estimator by numerical studies and simulations. We discuss the use of the constrained MINQU estimator in other cases in Section 4.

2. The Model Assumption and the Constrained MINQU Estimator

We consider a survey where interviewers are to be used. Suppose the enumeration areas can be conveniently grouped into P blocks. Randomly select t enumeration areas (clusters) out of each block, and randomly select p blocks out of the P blocks to carry the interpenetrated interview scheme. The t interviewers sent to the p blocks will split the workload of each enumeration area equally, hence each will carry n_{hi} interviews for the ith enumeration area of the hth block. The t interviewers sent to the remaining P-p blocks will each interview a whole enumeration area, thus each will carry tn_{hi} interviews.

Biemer and Stokes (1985) formulated the following linear model for the interpenetrated and noninterpenetrated assignments:

$$\begin{aligned} y^{(1)}_{hijk} &= \mu_{hi} + a_{hj} + e^{(1)}_{hijk}, \qquad h = 1, \ldots, p, \\ y^{(2)}_{hik} &= \mu_{hi} + a_{hi} + e^{(2)}_{hik}, \qquad h = p+1, \ldots, P, \end{aligned} \tag{2.1}$$

where $y^{(1)}_{hijk}$ and $y^{(2)}_{hik}$ are observations from the interpenetrated survey and the non-interpenetrated survey, respectively, μ_{hi} is the population mean in the hith EA, a_{hj} is the random effect of variance σ_a^2 associated with the jth interviewer in the hth block, and $e^{(1)}_{hijk}$ and $e^{(2)}_{hik}$ are the random residual errors of variances σ^2_{hi} associated with each unit in the p blocks and the P-p blocks, respectively. Biemer and Stokes (1985) showed that the synthesis MINQUE for the above model depends only on the interpenetrated data. Kleffe et al. (1991) indicate that this result holds for the general MINQUE.

To take advantage of the information in the noninterpenetrated blocks, Kleffe et al. (1991) made an assumption on the means of the clusters by treating them as random variables following the approach in Scott and Smith (1969). As their paper showed, this assumption can prove the unbiasedness of Fellegi's second estimator without the randomness assumption on the design of the survey

imposed in Biemer and Stokes (1985). Here we adopt the model assumption of Kleffe et al. (1991):

$$\begin{aligned} y^{(1)}_{hijk} &= \beta + b_h + c_{hi} + a_{hj} + e^{(1)}_{hijk}, \qquad h = 1, \ldots, p, \\ y^{(2)}_{hik} &= \beta + b_h + c_{hi} + a_{hi} + e^{(2)}_{hik}, \qquad h = p+1, \ldots, P, \end{aligned} \tag{2.2}$$

where $y^{(1)}_{hijk}$, $y^{(2)}_{hik}$, a_{hj}, a_{hi}, $e^{(1)}_{hijk}$ and $e^{(2)}_{hik}$ are defined as in the Biemer-Stokes model (2.1), β is the overall true mean for the clusters, b_h is the random effect of variance σ_b^2 associated with the hth block, c_{hi} is the random effect of variance σ_c^2 associated with the hith cluster.

The above model can also be written in matrix form as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{U}_1\xi_1 + \mathbf{U}_2\xi_2 + \mathbf{U}_3\xi_3 + \sum_{h=1}^{P}\sum_{i=1}^{t}\mathbf{U}_{hi}\xi_{hi}, \tag{2.3}$$

where $\mathbf{X}$ is the design vector for the mean parameter β, $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ and $\mathbf{U}_{hi}$ are the design matrices for the random effects ξ_1, ξ_2, ξ_3 and ξ_{hi}, respectively. (See Kleffe et al. (1991). Appendix A for the definitions of $\mathbf{U}$'s and ξ's.) A general method for estimating variance components from model (2.3) is the MINQUE of C. R. Rao. For model (2.3) the MINQU estimator of $\sigma^2 = (\sigma_b^2, \sigma_c^2, \sigma_a^2, \sigma_{hi}^2)'$ is obtained by solving the following system of linear equations:

$$\begin{aligned} &z_{11}\sigma_b^2 + z_{12}\sigma_c^2 + z_{13}\sigma_a^2 + \sum_{h=1}^{P}\sum_{i=1}^{t} z_{1hi}\sigma_{hi}^2 = q_1 \\ &z_{12}\sigma_b^2 + z_{22}\sigma_c^2 + z_{23}\sigma_a^2 + \sum_{h=1}^{P}\sum_{i=1}^{t} z_{2hi}\sigma_{hi}^2 = q_2 \\ &z_{13}\sigma_b^2 + z_{23}\sigma_c^2 + z_{33}\sigma_a^2 + \sum_{h=1}^{P}\sum_{i=1}^{t} z_{3hi}\sigma_{hi}^2 = q_3 \\ &z_{1hi}\sigma_b^2 + z_{2hi}\sigma_c^2 + z_{3hi}\sigma_a^2 + \sum_{h'=1}^{P}\sum_{i'=1}^{t} z_{h'i'hi}\sigma_{hi}^2 = q_{hi}, \\ &h = 1, \ldots, P, \quad i = 1, \ldots, t, \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} z_{ij} &= \|\mathbf{U}_i'\mathbf{R}\mathbf{U}_j\|^2, \quad z_{khi} = \|\mathbf{U}_k'\mathbf{R}\mathbf{U}_{hi}\|^2, \quad z_{hih'i'} = \|\mathbf{U}_{hi}'\mathbf{R}\mathbf{U}_{h'i'}\|^2, \\ q_i &= \|\mathbf{U}_i'\mathbf{R}\mathbf{y}\|^2, \qquad q_{hi} = \|\mathbf{U}_{hi}'\mathbf{R}\mathbf{y}\|^2 \\ \mathbf{R} &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}, \end{aligned}$$

$\mathbf{V}$ is the variance covariance matrix of $\mathbf{y}$,

$$\mathbf{V} = \sigma_b^2\mathbf{U}_1\mathbf{U}_1' + \sigma_c^2\mathbf{U}_2\mathbf{U}_2' + \sigma_a^2\mathbf{U}_3\mathbf{U}_3' + \sum_{hi}\sigma_{hi}^2\mathbf{U}_{hi}\mathbf{U}_{hi}',$$

and $\|\cdot\|^2$ denotes the Euclidean norm of a matrix.

The estimator obtained by solving equation (2.4) has locally minimum variance among all invariant unbiased estimators(LMIVQUE) under normality assumption of the data. However, there are several difficulties in solving equation (2.4) for σ^2. First, the number of linear equations $(Pt+3)$ gets so large with the increase in the number of blocks or the number of clusters within each block that solving the equations would become an infeasible task. Second, the unequal cluster size tn_{hi} will be reflected in the design matrices and in the variance covariance matrix $\mathbf{V}$, hence some convenient expressions for design matrices from balanced data, such as the Kronecker products of matrices, cannot be used. Third, the heterogeneity of the residual errors σ^2_{hi} will also be reflected through the variance covariance matrix and hence further complicate equation (2.4).

So far there are basically two ways to overcome the first and the third difficulties, namely, the synthesis MINQUE and the equal residual error assumption. The synthesis MINQUE, proposed by Hartley, Rao and LaMotte (1978), is the MINQUE obtained by using the identity matrix in the place of the variance covariance matrix in the MINQUE equation. This approach can indeed reduce the complexity of the MINQUE equations greatly. However, Rao and Kleffe (1988) and Kleffe et al. (1991) have shown by numerical comparisons that the synthesis MINQUE for the models they considered can be inefficient. The equal residual error assumption reduces the dimension of (2.4) from $Pt+3$ to 4. Although a dramatic reduction of the complexity, this approach may at the same time reduce the efficiency of the estimator. A study by Rao, Kaplan and Cochran (1981) considered the MINQUE obtained by assuming equal residual error (MINQUE EQUAL) for the one-way random effect model with unequal error variances, and found that (the variance of MINQUE EQUAL) "increases substantially when the implied assumption of the equality of the σ^2_i is not satisfied". Furthermore, both approaches will be deeply complicated by the unequal cluster size. In fact, all explicit estimators derived from the two approaches assumed equal cluster size.

Kleffe et al. (1991) give useful expressions for the design matrices from unbalanced data with general n_{hi} and σ^2_{hi}, and they also derive an expression for the matrix $\mathbf{R}$ by ingeniously avoiding the inversion of the matrix $\mathbf{V}$ which is of order $\sum_{hi} tn_{hi}$. While they acknowledge the difficulty in solving (2.4), they point out different assumptions on the residual errors to reduce the number of σ^2_{hi}, for example, $\sigma^2_{hi} = \sigma^2_h$. The explicit estimator $\hat{\sigma}^2_a$ they derive, however, is for the balanced case $n_{hi} = n$ and $\sigma^2_{hi} = \sigma^2_e$, which reduces the dimension of the MINQU equation from $Pt+3$ to 4.

In this paper we propose an approach to derive an efficient estimator for unbalanced data, i.e. unequal cluster sizes and heterogeneous residual errors, which are commonly encountered in complex surveys, and we find a way to reduce the large numbers of equations in the MINQUE equation to a manageable level.

Swallow and Searle (1978) in a numerical study on MINQUE from unbalanced data demonstrate that the ANOVA estimator of σ_e^2, obtained by equating quadratic forms to their expectations, proves to be an excellent estimator in the sense that its variance is very close to that of MINQUE. Therefore, there is strong evidence suggesting that we can estimate the σ_{hi}^2's efficiently by ANOVA.

Using the obtained unbiased estimators for the σ_{hi}^2's, we can move the $\hat{\sigma}_{hi}^2$'s to the right hand side of the equations, and hence we only need to solve:

$$\begin{aligned} z_{11}\sigma_b^2 + z_{12}\sigma_c^2 + z_{13}\sigma_a^2 &= q_1 - \sum_{h=1}^{P}\sum_{i=1}^{t} z_{1hi}\hat{\sigma}_{hi}^2 \\ z_{12}\sigma_b^2 + z_{22}\sigma_c^2 + z_{23}\sigma_a^2 &= q_2 - \sum_{h=1}^{P}\sum_{i=1}^{t} z_{2hi}\hat{\sigma}_{hi}^2 \\ z_{13}\sigma_b^2 + z_{23}\sigma_c^2 + z_{33}\sigma_a^2 &= q_3 - \sum_{h=1}^{P}\sum_{i=1}^{t} z_{3hi}\hat{\sigma}_{hi}^2. \end{aligned} \tag{2.5}$$

Let $\mathbf{Z} = (z_{ij})$ be the 3×3 matrix containing the z_{ij}'s, and let

$$\begin{aligned} a_1 &= \frac{Z_{31}}{|\mathbf{Z}|} = \frac{1}{|\mathbf{Z}|}(z_{12}z_{23} - z_{13}z_{22}), \\ a_2 &= \frac{Z_{32}}{|\mathbf{Z}|} = \frac{1}{|\mathbf{Z}|}(z_{12}z_{13} - z_{11}z_{23}), \\ a_3 &= \frac{Z_{33}}{|\mathbf{Z}|} = \frac{1}{|\mathbf{Z}|}(z_{11}z_{22} - z_{12}^2), \end{aligned}$$

where Z_{ij} is the cofactor of z_{ij}, and $|\mathbf{Z}|$ is the determinant of the matrix $\mathbf{Z}$,

$$|\mathbf{Z}| = z_{11}(z_{22}z_{33} - z_{23}^2) - z_{12}(z_{12}z_{33} - z_{13}z_{23}) + z_{13}(z_{12}z_{23} - z_{13}z_{22}).$$

Let $w_{hi} = a_1 z_{1hi} + a_2 z_{2hi} + a_3 z_{3hi}$; then the constrained MINQU estimator of σ_a^2 is

$$\hat{\sigma}_a^2 = a_1 q_1 + a_2 q_2 + a_3 q_3 - \sum_{h=1}^{P}\sum_{i=1}^{t} w_{hi}\hat{\sigma}_{hi}^2. \tag{2.6}$$

It can be seen from (2.5) that the number of blocks and the number of clusters have virtually no effect on the number of equations contained in (2.5). The estimator $\hat{\sigma}_a^2$ obtained from (2.6) remains unbiased provided that $\hat{\sigma}_{hi}^2$'s are unbiased for σ_{hi}^2's.

One ANOVA estimator of the σ_{hi}^2 may be:

$$\begin{aligned} \hat{\sigma}_{hi}^2 &= \frac{1}{t(n_{hi}-1)}\sum_{jk}(y_{hijk}^{(1)} - \bar{y}_{hi..}^{(1)} - \bar{y}_{h.j.}^{(1)} + \bar{y}_{h...}^{(1)})^2, \quad 1 \le h \le p, \\ \hat{\sigma}_{hi}^2 &= \frac{1}{tn_{hi}-1}\sum_{k}(y_{hik}^{(2)} - \bar{y}_{hi.}^{(2)})^2, \quad p+1 \le h \le P. \end{aligned} \tag{2.7}$$

This estimator of σ^2_{hi} assumes that there is no interaction between the clusters and the interviewers for the p interpenetrated blocks.

Another ANOVA estimator of σ^2_{hi} may be more robust when there are interactions in model (2.2):

$$\tilde{\sigma}^2_{hi} = \frac{1}{t(n_{hi}-1)} \sum_{jk} (y^{(1)}_{hijk} - \bar{y}^{(1)}_{hij.})^2, \qquad 1 \le h \le p,$$

$$\tilde{\sigma}^2_{hi} = \frac{1}{tn_{hi}-1} \sum_{k} (y^{(2)}_{hik} - \bar{y}^{(2)}_{hi.})^2, \quad p+1 \le h \le P. \tag{2.8}$$

In the following section we compare the efficiency of the constrained MINQU estimator to the full MINQU estimator for balanced and unbalanced data, respectively.

3. The Efficiency of the Constrained MINQU Estimator

3.1. The balanced case

The optimal MINQU estimator was derived for the balanced case where $n_{hi} = n$ and $\sigma^2_{hi} = \sigma^2_e$ by Kleffe et al. (1991), assuming a normal distribution of the data. The constrained MINQUE approach in the previous section reduces the dimension of the MINQUE equations by one from four to three in this case, hence it is not a significant reduction. Nevertheless , it is useful to compare the constrained MINQU estimator in the balanced case to the optimal estimator and see how much efficiency we have to compromise by reducing the dimension of the equations.

With $n_{hi} = n$ and $\sigma^2_{hi} = \sigma^2_e$, leting

$$c = \frac{tn}{1+tn\alpha_2},\ b = \frac{tn}{1+tn(\alpha_2+\alpha_3)},\ d = \frac{tn}{1+tn\alpha_3},\ a = \frac{tn}{1+tn(\alpha_2+\alpha_3+t\alpha_1)},$$

the constrained MINQU estimator can be written as:

$$\hat{\sigma}^2_a = b_2 Q_2 + b_3 Q_3 + b_4 Q_4 + b_e \hat{\sigma}^2_e, \tag{3.1}$$

where Q_2, Q_3 and Q_4 are defined in the same way as Kleffe et al. (1991):

$$Q_2 = \sum_{hi} (\bar{y}^{(1)}_{hi..} - \bar{y}^{(1)}_{h...})^2,\ Q_3 = \sum_{hi} (\bar{y}^{(2)}_{hi.} - \bar{y}^{(2)}_{h..})^2,\ Q_4 = \sum_{hj} (\bar{y}^{(1)}_{h.j.} - \bar{y}^{(1)}_{h...})^2,$$

and

$$b_2 = -\frac{b^2c^2(P-p)}{p(t-1)W}, \qquad b_3 = \frac{b^2c^2}{(t-1)W},$$

$$b_4 = \frac{c^2d^2p + d^2b^2(P-p)}{p(t-1)W}, \qquad b_e = -\frac{c^2d^2p + d^2b^2(P-p)}{ntW},$$

and $W = c^2d^2p + d^2b^2(P-p) + b^2c^2(P-p)$.

If we choose the non-interaction estimator of σ_e^2 of (2.7):

$$\hat{\sigma}_e^2 = \frac{1}{tT}\Big[\sum_{hijk}(y_{hijk}^{(1)} - \bar{y}_{hi..}^{(1)} - \bar{y}_{h.j.}^{(1)} + \bar{y}_{h...}^{(1)})^2 + \sum_{hik}(y_{hik}^{(2)} - \bar{y}_{hi.}^{(2)})^2\Big], \tag{3.2}$$

where $T = 1/(pt(n-1) + (P-p)(tn-1))$, $\hat{\sigma}_e^2$ is unbiased and if we let $Q_7 = \sum_{hijk}(y_{hijk}^{(1)} - \bar{y}_{hi..}^{(1)} - \bar{y}_{h.j.}^{(1)} + \bar{y}_{h...}^{(1)})^2 + \sum_{hik}(y_{hik}^{(2)} - \bar{y}_{hi.}^{(2)})^2$, then $\hat{\sigma}_a^2$ in (3.1) can be rewritten as:

$$\hat{\sigma}_a^2 = b_2Q_2 + b_3Q_3 + b_4Q_4 + b_7Q_7, \tag{3.3}$$

where $b_7 = b_e/tT$.

The optimal MINQU estimator given by Kleffe et al. for the balanced case is:

$$\hat{\sigma}_a^2(KPR) = a_2Q_2 + a_3Q_3 + a_4Q_4 + a_5Q_5, \tag{3.4}$$

where $Q_5 = \sum_{hijk}(y_{hijk}^{(1)} - \bar{y}_{h.j.}^{(1)})^2 + \sum_{hik}(y_{hik}^{(2)} - \bar{y}_{hi.}^{(2)})^2$, and a_2, a_3, a_4 and a_5 are given in their paper.

Kleffe et al. give the variances of the Q statistics used in their estimator. We further obtain that $V(Q_7) = 2tT\sigma_e^4$, and find that Q_7 is uncorrelated with Q_2, Q_3 and Q_4, and that the Q_5 used in Kleffe et al. is correlated with Q_2.

Kleffe and Rao (1993) derive conditions for the admissibility of the variance components estimators for model (2.2) in the set of quadratic unbiased estimators in the balanced case. An estimator $\hat{\theta}$ is admissible in a set of unbiased estimators if there is no other unbiased estimator in the class that is uniformly better than $\hat{\theta}$ with respect to variance. Kleffe and Rao (1993) show that any unbiased quadratic estimator of σ_a^2 can be written in the form of $s_a^2 = s_3 - \gamma W_0$, where

$$s_3 = \frac{1}{p(t-1)}Q_4 - \frac{1}{tnT}Q_7,$$

and

$$W_0 = -\frac{nt}{p(t-1)}Q_2 + \frac{nt}{(P-p)(t-1)}Q_3 - \frac{nt}{p(t-1)}Q_4 + \frac{1}{T}Q_7.$$

Kleffe and Rao (1993) prove that s_a^2 is admissible iff $c_f \le \gamma \le 0$, where

$$c_f = -\frac{1}{nt}\frac{P-p}{P}\max\Big\{1, \Big(\frac{P-p}{P} + \frac{T}{T+p(t-1)}\Big)^{-1}\Big\}.$$

Note that the constrained MINQU estimator $\hat{\sigma}_a^2$ using the non-interaction estimator of σ_e^2 given by (3.3) can be written as:

$$\hat{\sigma}_a^2 = s_3 - \gamma^* W_0,$$

where

$$\gamma^* = -\frac{1}{nt}\frac{b^2c^2(P-p)}{W}.$$

Therefore the admissibility of $\hat{\sigma}_a^2$ depends on the comparisons of γ^* with c_f. Since γ^* is a complicated function of the variance components, we made a numerical evaluation of γ^* at $n = 3$, $t = 2$, $P = 10$, $p = 4$ and $\sigma_e^2 = 1$ and plotted the values of γ^* at $\sigma_c^2 = 0.1, 1$ and 10 in the range of $0 \leq \sigma_a^2 \leq 10$ in Figure 1. Note in this situation $c_f = -0.1$. It can be seen from Figure 1 that γ^* showed a monotone decreasing trend and $c_f \leq \gamma^*$. Hence $\hat{\sigma}_a^2$ is admissible in these cases. Although the results are numerical for limited parameter values, they do demonstrate that the constrained MINQUE performs well relative to the full MINQUE in these cases. Although the MINQU estimator of Kleffe et al. is locally optimal, the admissibility of the constrained MINQU estimator indicates that the constrained MINQU is not uniformly worse than the full MINQU estimator.

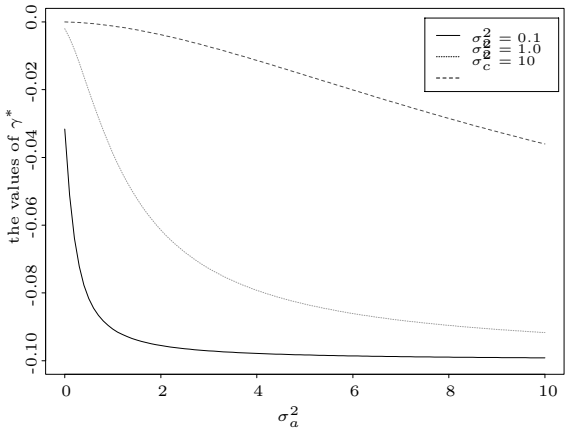

Figure 1. Numerical evaluations of γ^* for $n = 3$, $t = 2$, $P = 10$, $p = 4$ and $\sigma_e^2 = 1$

3.2. The unbalanced case

We carried out a simulation study to compare the constrained MINQU estimators to the MINQU estimators. Specifically, we compared the performances of four estimators: the full MINQU estimator, $\hat{\sigma}_a^2$(MINQUE), which is obtained by solving equation (2.4); the constrained MINQU estimator, $\hat{\sigma}_a^2$(CMINQUE), using non-interaction estimators in (2.7); the constrained MINQU estimator, $\tilde{\sigma}_a^2$(CMINQUE), using the robust estimator in (2.8); and the MINQU estimator by assuming equal random errors, i.e. $\sigma_{hi}^2 = \sigma^2$, $\hat{\sigma}_a^2$(EMINQUE). The EMINQU estimator is not necessarily unbiased. However, since this is the most widely employed approach in practice for unbalanced data, we included it for the purpose of comparison.

To demonstrate the impact of the unequal random errors on the estimation of the correlated response variance and also to simplify the configuration of the simulation, we assumed equal sample sizes for the clusters. We used P=10, p=4, t=2 and n=30. The parameters were: σ_b^2=0.1, σ_c^2=0.1, σ_{h1}^2=1.0, and σ_{h2}^2=4.0. One thousand simulations are performed for each parameter configuration.

We also considered the performances of the four estimators in the presence of model misspecification. Instead of using model (2.2), we generated random numbers under the following model for the interpenetrated blocks:

$$y_{hijk}^{(1)} = \beta + b_h + c_{hi} + a_{hj} + d_{hij} + e_{hijk}^{(1)},\ h = 1, \ldots, p, \tag{3.5}$$

where the random variable d_{hij} has a mean of zero and variance of σ_d^2. The non-interpenetrated blocks are assumed to follow the same model as in (2.2). The above model contains an interaction term between the clusters and the interviewers and the analysis is done assuming no interactions. In the simulation we set $\sigma_d^2 = 0.1$ for simulation numbers 6 to 9 and we increased the interaction term to $\sigma_d^2 = 0.5$ in simulation number 10 to dominate the other variance components.

Table 1. Monte Carlo values of the expected values of various estimators and their mean squared errors from the simulation. $P = 10$, $p = 4$, $t = 2$, $n = 30$. σ_b^2=0.1, σ_c^2=0.1. $\sigma_{h1}^2 = 1.0$ and $\sigma_{h2}^2 = 4.0$ for all simulations except No. 5 where equal random errors $\sigma_{hi}^2 = 1$ were used. $\sigma_d^2 = 0.1$ in the misspecified model (3.5) for simulations No. 6-9 and $\sigma_d^2 = 0.5$ for simulation No. 10.

Model	Simulation Number	σ_a^2	$\hat{\sigma}_a^2$(MINQUE)	$\hat{\sigma}_a^2$(CMINQUE)	$\tilde{\sigma}_a^2$(CMINQUE)	$\hat{\sigma}_a^2$(EMINQUE)
	1	0.02	0.0054	0.0497	0.0494	0.0091
			0.00085	0.00152	0.00150	0.000236
	2	0.1	0.0903	0.1036	0.1035	0.1155
			0.00997	0.00987	0.00987	0.00295
(2.1)	3	0.2	0.1952	0.1829	0.1829	0.2553
			0.03721	0.03743	0.03743	0.01094
	4	0.5	0.4998	0.4658	0.4660	0.6847
			0.18530	0.18617	0.18617	0.06544
	5	0.2	0.2010	0.2037	0.2037	0.2637
			0.02297	0.02296	0.02296	0.01067
	6	0.02	0.0494	0.0956	0.0946	0.0211
			0.00408	0.00882	0.00901	0.00013
	7	0.1	0.1054	0.1187	0.1184	0.1154
			0.01117	0.01150	0.01149	0.00299
(3.5)	8	0.2	0.1856	0.1719	0.1722	0.2453
			0.03400	0.03452	0.03451	0.01133
	9	0.5	0.4669	0.4305	0.4312	0.6618
			0.16772	0.17126	0.17115	0.0669
	10	0.5	0.3893	0.3478	0.3507	0.5883
			0.19410	0.20505	0.20489	0.07276

Table 1 presents the Monte Carlo values of the expected values of the various estimators and their mean squared errors. From the simulation results we can see that when the correlated response variance σ_a^2 is relatively large compared to the other variance components in the model, the two constrained MINQU estimators are close to the full MINQU estimator in terms of biases and the mean squared errors. The two constrained MINQU estimators are comparable in the situations considered here, although the constrained MINQU estimators using the robust random error estimator is slightly better than the constrained MINQUE with no interaction assumption under the misspecified model (3.5) with an interaction term. The EMINQU estimator has a rather large bias when σ_a^2 is large, although it has the smallest empirical mean squared error among the four estimators considered. No estimator is uniformly best.

The simulation results support our belief that using a robust estimator for the residual errors in the constrained MINQU equations does not compromise the efficiency of the other components of variance and is easy to compute.

4. Discussion

This paper deals with interpenetrated and noninterpenetrated assignments for stratified samples. For more complex survey designs the approach of Hartley and Rao (1978) allows the correlated response variance to be estimated for appropriate assignments.

Although the motivation for the constrained MINQU estimator given in Section 3 was to derive an efficient estimator of correlated response variance from unbalanced data in complex surveys, the approach can be used in other cases where reducing the complexity of the MINQUE equations is desired. Suppose we can obtain robust estimators on k out of K variance components, then instead of dealing with:

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_3 \\ \mathbf{A}_3' & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \end{bmatrix}, \tag{4.1}$$

assuming that θ_2 contains all the k variance components for which we have obtained estimates, we now only need to solve:

$$\mathbf{A}_1\theta_1 = \mathbf{q}_1 - \mathbf{A}_3\hat{\theta}_2, \tag{4.2}$$

which reduces the dimension of the MINQUE equations from K to $K - k$. The information contained in q_2 can be saved through an adequately chosen $\hat{\theta}_2$, hence no information is lost by the reduction of the equations.

Our approach can perhaps be better understood in the context of REML—the restricted maximum likelihood estimation of variance components, developed by Patterson and Thompson (1971). In theory, equating the first derivative of

the restricted likelihood function to zero results in the same estimating equations as MINQUE. However, while MINQUE adopts the use of prior values, REML obtains its estimates by iterative algorithms to ensure the maximum for its likelihood function. Our approach can be interpreted in REML as seeking the conditional maximum for the restricted likelihood function with some variance components known or fixed. In fact, the REML function installed in the statistical software Genstat is implemented with the option of fixing some variance components so that they would not need further iterations. However, since the survey model (2.2) does not fit into the general model assumption of having homogeneous residual errors, we can not derive our estimator by the REML function in Genstat.

In this paper we have not considered the impact of prior values on the constrained MINQU estimates. Various studies (Kleffe et al. (1991), Rao and Kleffe (1988)) have shown that the prior choice (0, 0) for $(\sigma_c^2/\sigma_e^2, \sigma_a^2/\sigma_e^2)$ is the least desirable. For the MINQU estimator to be less influenced by prior values, Rao and Kleffe (1988) suggest one-step iteration and they show in two examples that the estimators obtained by one-step iteration are practically independent of prior values. Since no gradient algorithm is imposed on our estimator we can not guarantee that continuous iteration will result in the convergence of the estimates.

References

Biemer, P. P. and Stokes, S. L. (1985). Optimal design of interviewer variance experiments in complex surveys. *J. Amer. Statist. Assoc.* **80**, 158-166.

Fellegi, I. P. (1974). An improved method of estimating the correlated response variance. *J. Amer. Statist. Assoc.* **69**, 496-501.

Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. (1961). Measurement errors in censuses and surveys. *Bull. Internat. Statist. Inst.* **38**, 359-374.

Hartley, H. O. and Rao, J. N. K. (1978). The estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurements* (Edited by N. K. Namboodiri), 35-43. Academic Press, New York.

Hartley, H. O., Rao, J. N. K. and LaMotte, L. (1978). A simple synthesis-based method of variance components estimation. *Biometrics* **34**, 233-242.

Kleffe, J., Prasad, N. G. N. and Rao, J. N. K. (1991). 'Optimal' estimation of correlated response variance under additive models. *J. Amer. Statist. Assoc.* **86**, 144-150.

Kleffe, J. and Rao, J. N. K. (1993). Inadmissibility but near optimality of an estimator of correlated response variance under additive models. *J. Statist. Plann. Inference* **36**, 151-164.

Lyberg, L. and Biemer, P. (1997). *Survey Measurement and Process Quality.* John Wiley, New York.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Rao, C. R. (1971). Estimation of variance and covariance components in linear models. *J. Multivariate Anal.* **1**, 257-275.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.* **67**, 112-115.

Rao, C. R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*. North-Holland.

Rao, P. S. R. S., Kaplan, J. and Cochran, W. G. (1981). Estimator for the one-way random effects model with unequal error variances. *J. Amer. Statist. Assoc.* **76**, 89-97.

Scott, A. J. and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *J. Amer. Statist. Assoc.* **64**, 830-840.

Swallow, W. H. and Searle, S. R. (1978). Minimum variance quadratic unbiased estimation (MIVQUE) of variance components. *Technometrics* **20**, 265-272.

Division of Biostatistics, Indiana University School of Medicine, 609 West Drive, RR135 Indianapolis, IN 46202-5119, U.S.A.

E-mail: sgao@mako.biostat.iupui.edu

Faculty of Mathematical Studies, Southampton University, Southampton, SO17 1BJ, U.K.

E-mail: tmfs@maths.soton.ac.uk

(Received February 1997; accepted October 1997)