J. R. Statist. Soc. B (1998)
60, *Part* 1, *pp.* 23–40

Weighting for unequal selection probabilities in multilevel models

D. Pfeffermann,

Hebrew University, Jerusalem, Israel

C. J. Skinner† and D. J. Holmes

University of Southampton, UK

and H. Goldstein and J. Rasbash

Institute of Education, London, UK

[*Read before* The Royal Statistical Society *at a meeting on the 'Design and analysis of complex sample surveys' organized by the* Research Section *on Wednesday, May 14th, 1997,* Dr D. Holt *in the Chair*]

Summary. When multilevel models are estimated from survey data derived using multistage sampling, unequal selection probabilities at any stage of sampling may induce bias in standard estimators, unless the sources of the unequal probabilities are fully controlled for in the covariates. This paper proposes alternative ways of weighting the estimation of a two-level model by using the reciprocals of the selection probabilities at each stage of sampling. Consistent estimators are obtained when both the sample number of level 2 units and the sample number of level 1 units within sampled level 2 units increase. Scaling of the weights is proposed to improve the properties of the estimators and to simplify computation. Variance estimators are also proposed. In a limited simulation study the scaled weighted estimators are found to perform well, although non-negligible bias starts to arise for informative designs when the sample number of level 1 units becomes small. The variance estimators perform extremely well. The procedures are illustrated using data from the survey of psychiatric morbidity.

Keywords: Hierarchical linear model; Iterative generalized least squares; Multistage sampling; Pseudolikelihood; Scaled weights; Variance components

1. Introduction

Sample surveys often employ multistage sampling schemes which involve unequal selection probabilities at some or all stages of the sampling process. Although these schemes are chosen mostly for cost and administrative reasons, the hierarchical population structure underlying such schemes is often of interest to survey data analysts. Multilevel models (Goldstein, 1995) provide an important class of regression models that may be employed to represent such structures.

Sampling schemes are commonly ignored in multilevel analyses of survey data. One argument in favour of this practice is that multilevel models can incorporate as covariates certain characteristics of the sampling design, such as strata and cluster indicators, and that conditionally on these characteristics the sampling design is ignorable in the sense of Rubin

†*Address for correspondence*: Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: cjs@soton.ac.uk

 1369–7412/98/60023

(1976). This argument may be inadequate, however, when units at any level of the hierarchy are selected with unequal probabilities in ways that are not accounted for by the model.

As an example, we consider the survey of psychiatric morbidity, conducted by the Office for National Statistics in 1993 with about 10000 adults living in private households in Great Britain (Meltzer *et al.*, 1995). The sample was obtained by a stratified multistage sampling design. Postal sectors on the 'small users' postcode address file were taken as primary sampling units. A sample of 200 of these sectors was selected by systematic probability proportional to size sampling. The size measure was the number of postal delivery points (corresponding approximately to addresses). Within each sampled sector, a simple random sample of 90 delivery points was selected. Interviewers visited the resulting $200 \times 90 = 18000$ delivery points and, among those containing at least one person aged 16–64 years, selected one such person at random. Thus the probabilities of selection of sectors and individuals vary according to sector size and delivery point size (number of eligible adults).

A multilevel analysis of data from such a survey, with individuals as level 1 units and postal sectors as level 2 units, may be of interest to assess the spatial homogeneity of psychiatric morbidity (e.g. Duncan *et al.* (1995)). It is conceivable that psychiatric variables of interest may be statistically related to either the sector size or the delivery point size. For example, the prevalence of neurotic symptoms and sector size might be positively associated through a common positive association with the sector's population density. Similarly, the prevalence of neurotic symptoms might be negatively associated with delivery point size because of the effect of lone parent households which tend to have higher levels of neurotic symptoms and lower average numbers of eligible adults. A data analyst may not, however, be given access to one or both of these size variables, for example for confidentiality reasons, or may not include them as covariates in the model if they are not scientifically meaningful.

When the sample selection probabilities are related to the response variable even after conditioning on covariates of interest, the conventional estimators of the model parameters may be (asymptotically) biased. The aim of this paper is to study weighting procedures that are appropriate for multilevel modelling, designed to correct for this bias. This corresponds to the analogous purpose of weighting in standard (single-level) regression models. For two reasons, weighting in multilevel models is not, however, a trivial extension of conventional methods of weighting.

Weighting in standard regression models can be viewed as an application of the 'pseudomaximum likelihood' (PML) approach as outlined in Skinner (1989), following ideas of Binder (1983). The basic idea of PML is that sample selection would not lead to bias if the values for all population units were observed, as in a census. If this were the case, we could compute the population (census) likelihood and achieve consistent estimation by maximizing this likelihood. When standard regression models are fitted to survey data, the finite population values are considered as independent so that the census log-likelihood is a sum which may be estimated consistently by simple weighting of the sample observations. The parameter value maximizing this estimated log-likelihood is the PML estimator which, under general conditions, is consistent for the corresponding model parameter. The first reason why multilevel models are different is that the finite population values are not independent in such models and so the census log-likelihood is not a simple finite population sum, implying that it cannot be estimated by simple weighting of the sample observations.

A second consequent reason why weighting for multilevel models is different in principle from conventional weighting is that the overall inclusion probabilities of the ultimate sample elements do not carry sufficient information for appropriate bias correction, unlike the single-level regression case. This fundamental issue will be illustrated in the following sections.

It should be emphasized that the multilevel model is assumed to be correctly specified and the weighting methods are designed solely to adjust for the effects of sampling that are not accounted for by the covariates included in the model. It is often argued that weighting can also protect against model misspecification (Pfeffermann, 1993) but this issue is not explored here.

Basic definitions and assumptions are set out in Section 2. The weighting approach is developed in Section 3 and its properties considered in Section 4. Scaling of the weights is discussed in Section 5 and variance estimation is considered in Section 6. The properties of the various estimators are evaluated in Section 7 by a simulation study and in Section 8 by analysing data from the survey of psychiatric morbidity. Section 9 contains some summarizing remarks.

Some proposals for weighting at the element level were made in Goldstein (1995). In their appendix, Pfeffermann and LaVange (1989) proposed a PML approach for the estimation of the fixed regression coefficients in a multilevel model. They also proposed consistent weighted estimators for the model variances, but these require knowledge of the joint second-order sample selection probabilities which are often not available. The present paper may be viewed as extending their work in various ways, in particular by also considering PML estimation of the variance components by using only first-order selection probabilities. Shah and LaVange (1994) also considered weighted estimation of the fixed regression coefficients. Longford (1995) and Graubard and Korn (1996) considered various weighted estimators of the variance components parameters for a simple two-level model.

2. Model and sampling design assumptions

Consider a two-level population, with M level 2 units (primary sampling units in survey sampling terminology) and N_j level 1 units within the jth level 2 unit ($j = 1, \ldots, M$). Let y_{ij} be the value of the response variable associated with the ith level 1 unit within the jth level 2 unit ($i = 1, \ldots, N_j; j = 1, \ldots, M$). Suppose that the y_{ij} are generated by the two-level model

$$y_{ij} = x_{ij}\beta + z_{ij}u_j + z_{0ij}v_{ij}, \tag{1}$$

where x_{ij}, z_{ij} and z_{0ij} are fixed covariate row vectors of dimensions p, q and 1 respectively, β is a fixed $p \times 1$ vector of parameters and u_j and v_{ij} are mutually independent normally distributed disturbances, $u_j \sim N(0, \Omega)$, $v_{ij} \sim N(0, \sigma^2)$. The term $z_{ij}u_j$ allows for random level 2 regression coefficients which, in the simplest case of $x_{ij1} \equiv 1$, $q = 1$ and $z_{ij} \equiv 1$ reduces to the random intercept model. It will commonly be the case that $z_{0ij} \equiv 1$, but the possibility of unequal z_{0ij} permits the representation of known patterns of heteroscedasticity within clusters. See Goldstein (1995) for further discussion of this model.

The following two-stage sampling scheme will be assumed. At the first stage, m level 2 units are selected with inclusion probabilities π_j ($j = 1, \ldots, M$). At the second stage, n_j level 1 units are selected within the jth selected level 2 unit with probabilities $\pi_{i|j}$. The (unconditional) sample inclusion probabilities are therefore $\pi_{ij} = \pi_{i|j}\pi_j$. The sampling mechanism may be informative in that the probabilities $\pi_{i|j}$ and π_j could be related to the error terms u_j and v_{ij} and hence to the y_{ij}.

3. Estimation

To apply PML estimation directly, we could in principle write down a closed form expression

for the 'census likelihood', estimate the log-likelihood function and then maximize the estimated function numerically. For computational efficiency and estimation simplicity we prefer, however, to begin with an established estimation method for the standard case, iterative generalized least squares (IGLS), and then adapt this by analogy with PML. The IGLS algorithm involves iterating between estimation of β and estimation of (Ω, σ^2) and is equivalent to maximum likelihood in the standard case under normality (Goldstein, 1986). We proceed by first writing down expressions for the 'census estimators' of β and (Ω, σ^2), which would be used in the IGLS algorithm if the entire population had been observed, and then replacing these census estimators by weighted sample estimators.

Consider therefore the IGLS algorithm for the hypothetical case where the values $(y_{ij}, x_{ij}, z_{ij}, z_{0ij})$ are observed for all population units, $i = 1, \ldots, N_j$, $j = 1, \ldots, M$. Let $Y_j = (y_{1j}, \ldots, y_{N_jj})'$, $X_j = (x_{1j}, \ldots, x_{N_jj})'$ and $e_j = (e_{1j}, \ldots, e_{N_jj})'$, where $e_{ij} = z_{ij}u_j + z_{0ij}v_{ij}$. Then the model defined by equation (1) for the population values may be expressed in matrix form as

$$Y_j = X_j\beta + e_j, \qquad e_j \sim N(0, V_j), \tag{2}$$

where $V_j = Z_j\Omega Z_j' + \sigma^2 D_j$, $Z_j = (z_{1j}', \ldots, z_{N_jj}')'$ and $D_j = \text{diag}(z_{01j}^2, \ldots, z_{0N_jj}^2)$.

Let $s = q(q+1)/2 + 1$ and let $\theta = (\theta_1, \ldots, \theta_s)'$ be the $s \times 1$ vector containing the distinct elements of Ω and $\theta_s = \sigma^2$. Then V_j may be expressed as a linear function of θ,

$$V_j(\theta) = \sum_{k=1}^{s} \theta_k G_{kj},$$

where $G_{kj} = Z_j H_{kj} Z_j' + \delta_{ks} D_j$, H_{kj} is a known $q \times q$ matrix containing 0s and 1s and δ_{ks} is the Kronecker delta. Let $E_{jj}[\beta] = (Y_j - X_j\beta)(Y_j - X_j\beta)'$ and note that $E_{jj}[\beta]$ has expectation $V_j(\theta)$. Following Anderson (1973) and Goldstein (1986), the IGLS algorithm involves the computation of a sequence of census estimators $\beta_C^{(r)}$ and $\theta_C^{(r)}$ of β and θ, $r = 1, 2, \ldots$, as follows.

Stage 1: set $\beta_C^{(r)} = P^{(r)-1}Q^{(r)}$, (3)

where $P^{(r)} = \Sigma_j X_j' V_{jr}^{-1} X_j$, $Q^{(r)} = \Sigma_j X_j' V_{jr}^{-1} Y_j$ and $V_{jr} = V_j(\hat{\theta}_C^{(r-1)})$, and Σ_j denotes sum over $j = 1, \ldots, M$.

Stage 2: set $\theta_C^{(r)} = R^{(r)-1}S^{(r)}$, (4)

where the klth element of the $s \times s$ matrix $R^{(r)}$ is $\Sigma_j \,\text{tr}(V_{jr}^{-1} G_{kj} V_{jr}^{-1} G_{lj})$ and the kth element of the $s \times 1$ vector $S^{(r)}$ is $\Sigma_j \,\text{tr}\{V_{jr}^{-1} G_{kj} V_{jr}^{-1} E_{jj}[\hat{\beta}^{(r)}]\}$. The iterative process is initialized at some value $\theta_C^{(0)}$. Under standard conditions, $\beta_C^{(r)}$ and $\theta_C^{(r)}$ converge to 'IGLS census estimators' β_C and θ_C as $r \to \infty$.

The census estimators are functions of the population values and hence are not operational if sampling is used. We therefore replace these census estimators by sample estimators $(\hat{P}^{(r)}, \hat{Q}^{(r)}, \hat{R}^{(r)}, \hat{S}^{(r)})$, with β and θ being estimated by the limiting values $\hat{\beta}$ and $\hat{\theta}$ of

$$\hat{\beta}^{(r)} = \hat{P}^{(r)-1}\hat{Q}^{(r)}, \qquad \hat{\theta}^{(r)} = \hat{R}^{(r)-1}\hat{S}^{(r)}, \qquad r = 1, 2, \ldots. \tag{5}$$

If $\hat{P}^{(r)}, \ldots, \hat{S}^{(r)}$ are taken as the sample versions of $P^{(r)}, \ldots, S^{(r)}$, $\hat{\beta}$ and $\hat{\theta}$ are the standard *unweighted IGLS estimators*. These estimators ignore the sampling scheme, however, and we therefore consider survey sampling methods to estimate consistently the finite population quantities $P^{(r)}, \ldots, S^{(r)}$.

Our proposed approach consists of replacing each sum over the level 2 population units j by

a sample sum weighted by $w_j = \pi_j^{-1}$ and each population sum over the level 1 units i by a sample sum weighted by $w_{i|j} = \pi_{i|j}^{-1}$ where π_j and $\pi_{i|j}$ are the corresponding selection probabilities. We refer to the resulting estimators as the *probability-weighted* IGLS (PWIGLS) *estimators*. If the w_j and $w_{i|j}$ are integers, the PWIGLS estimators could be obtained by duplicating $(y_{ij}, x_{ij}, z_{ij}, z_{0ij})$ $w_{i|j}$ times for each sample unit (i, j), duplicating the resulting sets of synthetic level 1 units w_j times for each j and then applying standard IGLS estimation. Such a procedure is, however, computationally very inefficient if the w_j or $w_{i|j}$ are large. Instead we seek a simpler approach.

We first obtain expressions for $P^{(r)}, \ldots, S^{(r)}$ as functions of sums over i and j. To simplify the exposition, we consider here the case $q = 1$, when there is just one random effect at level 2, and indicate the extension to $q > 1$ in Appendix A. When $q = 1$

$$\begin{aligned} P^{(r)} &= \sum_j T_{1j} - a_j T_{2j} T'_{2j}, \\ Q^{(r)} &= \sum_j T_{3j} - a_j T_{2j} T_{4j}, \end{aligned} \tag{6}$$

where $T_{1j} = \Sigma_i x_{ij} x'_{ij}/z_{0ij}^2$, $T_{2j} = \Sigma_i x_{ij} z_{ij}/z_{0ij}^2$, $T_{3j} = \Sigma_i x_{ij} y_{ij}/z_{0ij}^2$, $T_{4j} = \Sigma_i y_{ij} z_{ij}/z_{0ij}^2$, $a_j = (T_{5j} + \hat{\sigma}^2/\hat{\omega}^2)^{-1}$, $T_{5j} = \Sigma_i z_{ij}^2/z_{0ij}^2$, $\omega^2 = \mathrm{var}(u_j)$ is the scalar value of Ω, $\hat{\omega}^2$ and $\hat{\sigma}^2$ are the IGLS census estimators from iteration $r-1$ and Σ_i denotes sum over $i = 1, \ldots, N_j$. Similarly,

$$R^{(r)} = \begin{pmatrix} \sum_j b_j^2 & \sum_j b_j^2/T_{5j} \\ \sum_j b_j^2/T_{5j} & \sum_j \{\hat{\sigma}^{-4}(N_j - 1) + b_j^2/T_{5j}^2\} \end{pmatrix}, \qquad S^{(r)} = \begin{pmatrix} \sum_j b_j^2 \tilde{u}_j^2 \\ \sum_j (\hat{\sigma}^{-4} T_{6j} + b_j^2 \tilde{u}_j^2/T_{5j}) \end{pmatrix} \tag{7}$$

where $b_j = (\hat{\omega}^2 + \hat{\sigma}^2/T_{5j})^{-1}$, $T_{6j} = \Sigma_i \tilde{v}_{ij}^2$, $\tilde{u}_j = (\Sigma_i e_{ij} z_{ij}/z_{0ij}^2)/T_{5j}$, $\tilde{v}_{ij} = (e_{ij} - z_{ij}\tilde{u}_j)/z_{0ij}$ and $e_{ij} = y_{ij} - x_{ij}\hat{\beta}^{(r)}$. Note that, from equations (6) and (7), $P^{(r)}, \ldots, S^{(r)}$ are functions of sums over i and j, as desired.

The PWIGLS estimators are obtained by replacing population sums of the form $\Sigma_j d_j$ and $\Sigma_i d_{ij}$ by the corresponding sample sums $\Sigma_j^{\mathrm{s}} w_j d_j$ and $\Sigma_i^{\mathrm{s}} w_{i|j} d_{ij}$, where Σ_j^{s} denotes sum over the sample level 2 units j and Σ_i^{s} denotes sum over the sample level 1 units i. Note that the weighted sample sums are unbiased and consistent for the corresponding population sums under the randomization distribution induced by the sampling process (see Section 4). We estimate N_j in $R^{(r)}$ by $\hat{N}_j = \Sigma_i^{\mathrm{s}} w_{i|j}$, even if the N_j are known, since we found in our simulation study that the use of N_j leads to slightly more biased estimates of σ^2.

Since computer software for the standard IGLS algorithm is widely available, it would be attractive if the PWIGLS algorithm could be implemented by transforming the data and applying the standard IGLS algorithm to the transformed data. We therefore consider the following transformation.

Step A: replace z_{ij} by $w_j^{-1/2} z_{ij}$; replace z_{0ij} by $w_j^{-1/2} w_{i|j}^{-1/2} z_{0ij} = w_{ij}^{-1/2} z_{0ij}$.

Following the application of step A, it is straightforward to show that the sample versions of $P^{(r)}$ and $Q^{(r)}$ defined in equations (6) may be expressed as

$$\begin{aligned} \hat{P}^{(r)} &= \sum_j^{\mathrm{s}} w_j (\hat{T}_{1j} - \hat{a}_j \hat{T}_{2j} \hat{T}'_{2j}), \\ \hat{Q}^{(r)} &= \sum_j^{\mathrm{s}} w_j (\hat{T}_{3j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{4j}), \end{aligned} \tag{8}$$

where $\hat{T}_{1j} = \Sigma_i^s w_{i|j} x_{ij} x_{ij}'/z_{0ij}^2$, $\hat{T}_{2j} = \Sigma_i^s w_{i|j} x_{ij} z_{ij}/z_{0ij}^2$, $\hat{T}_{3j} = \Sigma_i^s w_{i|j} x_{ij} y_{ij}/z_{0ij}^2$, $\hat{T}_{4j} = \Sigma_i^s w_{i|j} y_{ij} z_{ij}/z_{0ij}^2$, $\hat{a}_j = (\hat{T}_{5j} + \hat{\sigma}^2/\hat{\omega}^2)^{-1}$ and $\hat{T}_{5j} = \Sigma_i^s w_{i|j} z_{ij}^2/z_{0ij}^2$. The estimators $\hat{P}^{(r)}$ and $\hat{Q}^{(r)}$ in equations (8) are precisely the PWIGLS estimators defined before and so stage 1 of the PWIGLS algorithm is achieved simply by transforming the data using step A and then applying stage 1 of the standard IGLS algorithm. For given $\hat{\theta} = (\hat{\omega}^2, \hat{\sigma}^2)'$, $\hat{\beta} = \hat{P}^{(r)-1}\hat{Q}^{(r)}$ is the same estimator as in Pfeffermann and LaVange (1989). It turns out that step A also achieves the necessary weighting for estimating $S^{(r)}$ in equations (7). For, following step A, the sample version of $S^{(r)}$ becomes

$$\hat{S}^{(r)} = \begin{pmatrix} \sum_j^s w_j \hat{b}_j^2 \hat{u}_j^2 \\ \sum_j^s w_j (\hat{\sigma}^{-4} \hat{T}_{6j} + \hat{b}_j^2 \hat{u}_j^2/\hat{T}_{5j}) \end{pmatrix} \tag{9}$$

where $\hat{b}_j = (\hat{\omega}^2 + \hat{\sigma}^2/\hat{T}_{5j})^{-1}$, $\hat{T}_{6j} = \Sigma_i^s w_{i|j} \hat{v}_{ij}^2$, $\hat{u}_j = (\Sigma_i^s w_{i|j} e_{ij} z_{ij}/z_{0ij}^2)/\hat{T}_{5j}$ and $\hat{v}_{ij} = (e_{ij} - z_{ij}\hat{u}_j)/z_{0ij}$. Unfortunately, the same is not true for estimating $R^{(r)}$, since application of only step A yields

$$\hat{R}_A^{(r)} = \begin{pmatrix} \sum_j^s \hat{b}_j^2 & \sum_j^s \hat{b}_j^2/\hat{T}_{5j} \\ \sum_j^s \hat{b}_j^2/\hat{T}_{5j} & \sum_j^s \{\hat{\sigma}^{-4}(n_j - 1) + \hat{b}_j^2/\hat{T}_{5j}^2\} \end{pmatrix} \tag{10}$$

which differs from the PWIGLS estimator:

$$\hat{R}^{(r)} = \begin{pmatrix} \sum_j^s w_j \hat{b}_j^2 & \sum_j^s w_j \hat{b}_j^2/\hat{T}_{5j} \\ \sum_j^s w_j \hat{b}_j^2/\hat{T}_{5j} & \sum_j^s w_j \{\hat{\sigma}^{-4}(\hat{N}_j - 1) + \hat{b}_j^2/\hat{T}_{5j}^2\} \end{pmatrix}. \tag{11}$$

We therefore propose to augment step A with the necessary additional adjustment to $\hat{R}_A^{(r)}$.

Step B: (a) insert the weights w_j into each of the sums Σ_j^s in $\hat{R}_A^{(r)}$;
(b) replace n_j in the (2, 2) element of $\hat{R}_A^{(r)}$ by $\hat{N}_j = \Sigma_j^s w_{i|j}$.

In summary, PWIGLS estimation may be implemented by first transforming the data by step A and then applying the standard IGLS algorithm, modified by step B. Initial values $\hat{\beta}^{(0)}$ and $\hat{\theta}^{(0)}$ for the PWIGLS algorithm may be computed as $\hat{\beta}^{(0)} = (\Sigma_j^s w_j \hat{T}_{1j})^{-1} \Sigma_j^s w_j \hat{T}_{3j}$, $\hat{\omega}^{(0)2} = 0$ and $\hat{\sigma}^{(0)2} = \Sigma_j^s w_j \hat{T}_{6j}^{(0)}/\Sigma_j^s w_j(\hat{N}_j - 1)$, where $\hat{T}_{6j}^{(0)} = \Sigma_i^s w_{i|j}(e_{ij}^{(0)} - z_{ij}\hat{u}_j^{(0)})^2/z_{0ij}^2$, $e_{ij}^{(0)} = y_{ij} - x_{ij}\hat{\beta}^{(0)}$ and $\hat{u}_j^{(0)} = \Sigma_i^s w_{i|j} e_{ij}^{(0)} z_{ij}/z_{0ij}^2/\hat{T}_{5j}$.

4. Consistency of probability-weighted iterative generalized least squares estimators

The PWIGLS estimators $\hat{\beta}^{(r)}$ and $\hat{\theta}^{(r)}$ defined in Section 3 are consistent for the corresponding census estimators $\hat{\beta}_C^{(r)}$ and $\theta_C^{(r)}$ under the randomization (repeated sampling) distribution, subject to the standard weak kinds of regularity conditions on the sampling scheme required for the consistency of Horvitz–Thompson-type estimators. Note that the establishment of randomization-based consistency properties requires a formulation of the way that the sample and population sizes mutually increase (Isaki and Fuller, 1982). In particular, to establish randomization-based consistency of the proposed PWIGLS approach requires both

m and the n_j to increase. This is because the sums over level 1 units enter non-linearly and the bias effect of this non-linearity may not vanish for small n_j. For example, $\hat{b}_j^2/\hat{T}_{5j}$ in equation (11) is biased for b_j^2/T_{5j} in equation (7) and this bias need not disappear in the weighted sum over j if n_j is fixed.

If the estimators $\hat{\beta}^{(r)}$ and $\hat{\theta}^{(r)}$ are consistent for the census estimators $\hat{\beta}_C^{(r)}$ and $\theta_C^{(r)}$, the limiting (as $r \to \infty$) PWIGLS estimators $\hat{\beta}$ and $\hat{\theta}$ will converge to the corresponding census IGLS estimators β_C and θ_C. Since the latter are consistent for the model parameters, the PWIGLS estimators are likewise consistent for these parameters with respect to the joint distribution induced by the model and the sampling scheme. See, for example, Pfeffermann (1993) for further discussion.

The requirement that both m and the n_j increase is unattractive since, in practice, the n_j are often small. In fact, consistency of $\hat{\beta}^{(r)}$ but not $\hat{\theta}^{(r)}$ can also be established when only m increases assuming fixed values $\hat{\sigma}^2$ and $\hat{\omega}^2$. To see this, rewrite equations (8) as the population sums $\hat{P}^{(r)} = \Sigma_j \Sigma_i k_{ij}^s x_{ij}$ and $\hat{Q}^{(r)} = \Sigma_j \Sigma_i k_{ij}^s y_{ij}$, where the k_{ij}^s depend on the selected samples ($k_{ij}^s = 0$ if ij is not sampled) and the values x_{ij}, z_{ij}, z_{0ij}, w_j and w_{ij} but not on the y_{ij}. It follows under standard regularity conditions on the k_{ij}^s that, as $m \to \infty$, $\hat{P}^{(r)-1}\{\Sigma_j \Sigma_i E_p[k_{ij}^s]x_{ij}\} \to I$ and $\{\Sigma_j \Sigma_i E_p[k_{ij}^s]x_{ij}\}^{-1}\hat{Q}^{(r)} \to \beta$ in probability, where E_p denotes expectation with respect to the randomization distribution. Hence $\hat{\beta}^{(r)}$ is consistent for β as m increases, given $\hat{\sigma}^2$ and $\hat{\omega}^2$.

5. Scaled estimators

In this section we consider scaling the weights in the PWIGLS estimators to reduce small sample biases, while retaining consistency. We note first from equations (8), (9) and (11) that $\hat{\beta}$ and $\hat{\theta}$ are invariant to scale multiplication of the w_j. Hence we restrict attention to scaling the $w_{i|j}$, i.e. replacing each $w_{i|j}$ in the expressions for the PWIGLS estimators by $w_{i|j}^* = \lambda_j w_{i|j}$, where the λ_j are constants to be determined. We write the resulting estimators as $\hat{\beta}(\lambda_j)$ and $\hat{\theta}(\lambda_j)$.

In choosing the scaling factors we shall treat m and the N_j as large, a common situation, but treat the n_j as fixed and possibly small. The argument presented in Section 4 for the consistency of $\hat{\beta}^{(r)}$ as $m \to \infty$ for fixed n_j is equally valid when the $w_{i|j}$ are scaled, provided that the λ_j do not depend on the y_{ij}. This suggests that the choice of the λ_j may not have a large effect on the bias of $\hat{\beta}(\lambda_j)$ when m is large. Hence, we focus on choosing λ_j to reduce the bias of the estimator $\hat{\theta}(\lambda_j)$ of the variance components. To determine a simple expression for the preferred λ_j we make some approximations. First, we consider asymptotic expressions for $\hat{\omega}^{2(r)}(\lambda_j)$ and $\hat{\sigma}^{2(r)}(\lambda_j)$, defined by equations (5), (9) and (11), where $\hat{N}_j$, $\hat{T}_{5j}$ and $\hat{T}_{6j}$ increase in proportion to N_j, say, and then omit terms of lower order, to obtain

$$\hat{\omega}^{2(r)}(\lambda_j) \doteq \left[\sum_j^s w_j \hat{b}_j^2(\lambda_j)\{\hat{u}_j^2 - \hat{\sigma}^{2(r)}(\lambda_j)/\lambda_j \hat{T}_{5j}\}\right] \Big/ \sum_j^s w_j \hat{b}_j^2(\lambda_j), \tag{12}$$

$$\hat{\sigma}^{2(r)}(\lambda_j) \doteq \sum_j^s w_j \lambda_j \hat{T}_{6j} \Big/ \sum_j^s w_j(\lambda_j \hat{N}_j - 1), \tag{13}$$

where $\hat{b}_j(\lambda_j)$ denotes the value of $\hat{b}_j$ when $w_{i|j}$ is replaced by $\lambda_j w_{i|j}$ and so forth. Next, we evaluate the expectation E_ξ with respect to the model by assuming that sampling of level 1 units (but not level 2 units) is approximately non-informative, which we expect to be the case in most practical applications. Noting that

$$\hat{u}_j \doteq u_j + \frac{\sum_i^s w_{i|j} v_{ij} z_{ij}/z_{0ij}}{\hat{T}_{5j}}$$

and treating $\hat{\omega}^2$ and $\hat{\sigma}^2$ in $\hat{b}_j(\lambda_j)$ as fixed, it follows from expression (12) that for sufficiently large m

$$E_\xi[\hat{\omega}^{2(r)}(\lambda_j)] - \omega^2 \doteq \sum_j^s \left\{ w_j \hat{b}_j^2(\lambda_j) \frac{\sigma^2/\tilde{\lambda}_j - E_\xi[\hat{\sigma}^{2(r)}(\lambda_j)]/\lambda_j}{\hat{T}_{5j}} \right\} \Big/ \sum_j^s w_j \hat{b}_j^2(\lambda_j),$$

where $\tilde{\lambda}_j = \hat{T}_{5j}/\hat{T}_{7j}$ and $\hat{T}_{7j} = \Sigma_i^s w_{i|j}^2 z_{ij}^2/z_{0ij}^2$. Also $E_\xi[\hat{T}_{6j}] \doteq \sigma^2(\hat{N}_j - \hat{T}_{7j}/\hat{T}_{5j})$, and so from expression (13)

$$E_\xi[\hat{\sigma}^{2(r)}(\lambda_j)] - \sigma^2 \doteq \sum_j^s w_j(1 - \lambda_j/\tilde{\lambda}_j)\sigma^2 \Big/ \sum_j^s w_j(\lambda_j \hat{N}_j - 1).$$

Both these expressions for bias tend to 0 as $\hat{N}_j$ and $\hat{T}_{5j}$ increase for fixed λ_j, illustrating that scaling the weights $w_{i|j}$ does not affect the asymptotic model unbiasedness of the PWIGLS estimator of θ even with fixed n_j. Note also that these two expressions, representing the $O(N_j^{-1})$ terms in the bias, are 0 when $\lambda_j = \tilde{\lambda}_j$. This suggests that we take $\tilde{\lambda}_j$ as our choice of λ_j to reduce the bias of $\hat{\theta}$. However, $\tilde{\lambda}_j$ depends on the z_{ij} and z_{0ij} and this would become complicated when models with several choices of z_{ij} or z_{0ij} are entertained. As a further simplification we suppose therefore that the z_{ij} and z_{0ij} are approximately uncorrelated with the $w_{i|j}$ within level 2 units so that $\tilde{\lambda}_j$ becomes approximately $\tilde{w}_j^{-1}$, where $\tilde{w}_j = \Sigma_i^s w_{i|j}^2/\Sigma_i^s w_{i|j}$. (In fact, $\tilde{\lambda}_j = \tilde{w}_j^{-1}$ for the random intercept model where $z_{ij} \equiv z_{0ij} \equiv 1$). We refer to the scaled weight $w_{i|j}^* = w_{i|j}/\tilde{w}_j$ as *scaling method 1*. The $\tilde{w}_j$ may be interpreted as the 'design effect' required to reduce the 'naïve sample size' $\hat{N}_j$ in the unscaled PWIGLS estimator to the 'effective sample size' $(\Sigma_i^s w_{i|j})^2/\Sigma_i^s w_{i|j}^2$.

As an alternative scaling method 2, we consider $\lambda_j = \bar{w}_j^{-1}$, where $\bar{w}_j = \Sigma_i^s w_{i|j}/n_j$. This factor reduces the naïve sample size $\hat{N}_j$ to the actual sample size n_j and, being similar to $\tilde{w}_j$, might also be expected to reduce the bias of $\hat{\theta}$ when the n_j are not large. It has two additional advantages. First it avoids the need for part (b) of step B since the scaled version of $\hat{N}_j$ becomes identical with n_j. Second, for the random intercept model with $q = 1$, $z_{ij} \equiv z_{0ij} \equiv 1$, and equal sample sizes n_j, step B is made redundant altogether provided that the w_j are also scaled to sum to m. This is so since $\hat{N}_j$, $\hat{b}_j$ and $\hat{T}_{5j}$ are constant under scaling and the incorporation of the weights w_j in the sums Σ_j^s in $\hat{R}_A^{(r)}$ in equation (10) is redundant. Note that selection of the level 2 units with unequal probabilities and equal-sized samples of level 1 units is quite common in practice.

Finally we note that if the w_j and $w_{i|j}$ are constant across i and j then for both scaling methods the scaled PWIGLS estimator is identical with the standard IGLS estimator unlike the unscaled estimator. In this case the sampling can be assumed to be non-informative and so the scaled estimators should be asymptotically efficient.

6. Variance estimation

We consider estimating the variance of the PWIGLS estimators with respect to the combined model and randomization distributions. It follows from Pfeffermann (1993) that for sufficiently small sampling fractions at both levels this variance can be estimated consistently by estimating just the randomization variance. This can be implemented by use of the delta

method, which becomes particularly simple if the level 2 units are treated as being selected with replacement, permitting us to consider only the level 2 selection in the computation of the variance estimators (Skinner, 1989). For small fractions m/M, this is generally not a restrictive assumption. In what follows we give the variance formulae for the unscaled estimator for the case $q = 1$. Estimation of the variances of the scaled estimators or the unweighted estimators is carried out in the same way. For the case $q = 1$, the delta method variance estimator of $\hat{\beta}$ is

$$v(\hat{\beta}) = \hat{P}^{-1}\left(\frac{m}{m-1}\right)\left(\sum_j^s w_j^2 c_j c_j'\right)\hat{P}^{-1} \tag{14}$$

where $\hat{P} = \lim_{r\to\infty}(\hat{P}^{(r)})$, $c_j = \Sigma_i^s w_{i|j} x_{ij} e_{ij}/z_{0ij}^2 - \hat{a}_j \hat{T}_{2j} \Sigma_i^s w_{i|j} e_{ij} z_{ij}/z_{0ij}^2$ and $e_{ij} = y_{ij} - x_{ij}\hat{\beta}$.

Similarly, the delta method variance estimator of $\hat{\theta}$ is

$$v(\hat{\theta}) = \hat{R}^{-1}\left(\frac{m}{m-1}\right)\left(\sum_j^s w_j^2 d_j d_j'\right)\hat{R}^{-1} \tag{15}$$

where $\hat{R} = \lim_{r\to\infty}(\hat{R}^{(r)})$ and

$$d_j = \begin{pmatrix} \hat{b}_j^2(\hat{u}_j^2 - \hat{\omega}^2 - \hat{\sigma}^2/\hat{T}_{5j}) \\ \hat{\sigma}^{-4}\{\hat{T}_{6j} - (\hat{N}_j - 1)\hat{\sigma}^2\} + \dfrac{\hat{b}_j^2(\hat{u}_j^2 - \hat{\omega}^2 - \hat{\sigma}^2/\hat{T}_{5j})}{\hat{T}_{5j}} \end{pmatrix}.$$

7. Simulation study

7.1. Design of experiment

To evaluate the properties of the various estimators, we conducted a small simulation study. Finite population values y_{ij} were generated from the model $y_{ij} = \beta + u_j + v_{ij}$; $u_j \sim N(0, \omega^2)$, $v_{ij} \sim N(0, \sigma^2)$, $j = 1, \ldots, M$, $i = 1, \ldots, N_j$. Results for the values $\beta = 1$, $\omega^2 = 0.2$ and $\sigma^2 = 0.5$ are reported here. The number of level 2 units in the population was $M = 300$. The sizes N_j were determined by $N_j = 75\ \exp(\tilde{u}_j)$, with $\tilde{u}_j$ generated from $N(0, \omega^2)$, truncated below by -1.5ω and above by 1.5ω. For $\omega^2 = 0.2$ the N_j lie in the range [38, 147], with mean around 80. We report results for the following sampling schemes.

(a) *Informative at both levels*: m level 2 units were sampled with probability proportional to a 'measure of size' X_j, so that $\pi_j = mX_j/\Sigma_1^M X_j$; the measure X_j was determined in the same way as N_j but with $\tilde{u}_j$ replaced by u_j, the random effect at level 2. The level 1 units in the jth sampled level 2 unit were partitioned into two strata according to whether $v_{ij} > 0$ or $v_{ij} \leqslant 0$ and simple random samples of sizes $0/25n_j$ and $0.75n_j$ were selected from the respective strata. The sizes n_j were either fixed, $n_j = n_0$, or proportional to N_j.

(b) *Informative only at level 2*: the scheme is the same as (a), except that simple random sampling was employed for the selection of level 1 units within each sampled level 2 unit.

(c) *Non-informative*: the scheme is the same as (b), except that the size measure X_j was set equal to N_j.

For each sampling scheme and parameter values the process of generating the finite population values and selecting the sample (one sample per population) was repeated 1000 times. For each sample the standard (unweighted) IGLS estimators and the PWIGLS

estimators (unscaled and two scaled versions) as well as their corresponding variance estimators were computed. To assess the importance of step B of the weighting process, the scaled (method 2) estimator obtained by application of only step A was also computed. Application of only step A without scaling yields absurd results for $(\hat{\omega}^2, \hat{\sigma}^2)$ since these estimators solve the equations $\hat{R}_A^{(r)}\hat{\theta} = \hat{S}^{(r)}$ with the coefficients in $\hat{R}_A^{(r)}$ being unweighted (equation (10)) and $\hat{S}^{(r)}$ being weighted (equation (9)).

7.2. Results

The results were generally more sensitive to the sample numbers of level 1 units than to the sample number of level 2 units. Hence, we report only results for the case where the sample number of level 2 units is $m = 35$. Increasing this value to $m = 75$ was generally found not to affect biases greatly, *ceteris paribus*. We report results for four different sample sizes within level 2 units: a fixed sample size $n_j = n_0 = 38$; proportional allocation $n_j = 0.4N_j$, for which the mean of the n_j is about 38; a fixed size $n_j = n_0 = 9$; proportional allocation $n_j = 0.1N_j$ (mean of about 9).

Tables 1–3 show the simulation means of the various estimators. It is evident that the unweighted estimators of each parameter can be seriously biased when the sampling at both levels is informative. When the sampling is only informative at level 2, the bias in the estimation of σ^2, a within level 2 unit parameter, disappears, but the unweighted estimators of β and ω^2 remain biased. The bias largely disappears when the design is non-informative. The minor bias in the estimation of ω^2 appears to represent the usual small sample bias of maximum likelihood estimation.

The unscaled weighted estimator performs well in removing the bias of the unweighted estimator for the larger sample sizes ($n_j = 38$ or $n_j = 0.4N_j$). This is evident under both informative sampling schemes with all three parameters. For the smaller sample sizes ($n_j = 9$

Table 1. Simulation means of point estimators of β†

Sampling design	*Unweighted estimator*	*Weighted estimators*			
		Unscaled	*Scaled*		*Step A only*
			1	*2*	
Informative at both levels					
$n_j = 38$	1.41	1.00	1.00	1.00	1.00
$n_j = 0.4N_j$	1.46	1.00	1.00	1.00	1.00
$n_j = 9$	1.48	1.00	1.00	1.00	1.00
$n_j = 0.1N_j$	1.51	1.04	1.04	1.03	1.03
Informative only at level 2					
$n_j = 38$	1.17	1.01	1.01		1.01
$n_j = 0.4N_j$	1.17	1.01	1.01		1.01
$n_j = 9$	1.17	1.01	1.01		1.01
$n_j = 0.1N_j$	1.17	1.00	1.01		1.00
Non-informative					
$n_j = 38$	1.00	1.00	1.00		1.00
$n_j = 0.4N_j$	0.99	0.99	0.99		0.99
$n_j = 9$	1.00	1.00	1.00		1.00
$n_j = 0.1N_j$	1.00	1.00	1.00		1.00

†The true value of β is 1; the number of sampled level 2 units is $m = 35$; the number of replications is 1000.

Table 2. Simulation means of point estimators of ω^2†

Sampling design	*Unweighted estimator*	*Weighted estimators*			
		Unscaled	*Scaled*		*Step A only*
			1	*2*	
Informative at both levels					
$n_j = 38$	0.191	0.197	0.188	0.191	0.191
$n_j = 0.4N_j$	0.178	0.201	0.181	0.189	0.189
$n_j = 9$	0.158	0.220	0.137	0.169	0.169
$n_j = 0.1N_j$	0.155	0.252	0.131	0.173	0.174
Informative only at level 2					
$n_j = 38$	0.183	0.196	0.190		0.190
$n_j = 0.4N_j$	0.182	0.201	0.189		0.189
$n_j = 9$	0.179	0.235	0.185		0.185
$n_j = 0.1N_j$	0.181	0.261	0.189		0.189
Non-informative					
$n_j = 38$	0.193	0.198	0.192		0.192
$n_j = 0.4N_j$	0.194	0.205	0.194		0.195
$n_j = 9$	0.194	0.242	0.192		0.192
$n_j = 0.1N_j$	0.189	0.259	0.190		0.191

†The true value of ω^2 is 0.2; the number of sampled level 2 units is $m = 35$; the number of replications is 1000.

or $n_j = 0.1N_j$), the bias in the estimation of ω^2 and σ^2 remains non-negligible, however, and of similar magnitude and direction under all three sampling schemes. Other simulations not reported here with $m = 75$ yielded similar biases. It appears that the sample sizes within level 2 units is the critical factor affecting the bias of the unscaled PWIGLS estimators.

Next, we discuss the performance of the scaled estimators. As suggested in Section 5, scaling leaves the estimator $\hat{\beta}$ in Table 1 approximately unbiased. The theory in Section 5 suggests the use of scaling method 1 to reduce bias in the estimation of ω^2 and σ^2 for non-informative sampling at level 1. For sampling schemes (b) and (c) the two scaled estimators are identical and, allowing for the standard small sample bias of the maximum likelihood estimator, scaling acts to reduce the bias of both the unweighted estimator and the unscaled weighted estimator in the case of small sample sizes. For the informative sampling scheme (a), method 1 seems to overcorrect and scaling method 2 is preferable, although it still displays non-negligible bias for the small sample sizes. The bias reduction from scaling is even more evident in Table 3 in the estimation of σ^2, although again method 1 seems to over-correct and some bias arises for method 2 for the smaller sample sizes in scheme (a).

The use of only step A for the scaled estimator yields very similar results for scaling method 2 in most cases. (As noted in Section 5, when the n_j are fixed, the two estimators are identical.) The only exception is the estimation of σ^2 under sampling scheme (c) with varying sample sizes n_j. In this case n_j is related to w_j as both n_j and w_j^{-1} are proportional to N_j. This bias appears to arise because the absence of the weights w_j in the (2, 2) element of $\hat{R}_A^{(r)}$ in equation (10) implied by the use of only step A leads to bias if w_j is related to n_j.

Table 4 contains results for the standard deviations of the point estimators and for the means of the sample estimators of these standard errors. The relative properties of the various estimators for the smaller sample sizes ($n_j = 9$ and $n_j = 0.1N_j$) were similar to those for the larger sample sizes ($n_j = 38$ and $n_j = 0.4N_j$) and so only the latter results are reported here. As expected, weighting leads to some inflation of standard errors, but for the cases of the larger

Table 3. Simulation means of point estimators of σ^2†

Sampling design	*Unweighted estimator*	*Weighted estimators*			
		Unscaled	*Scaled*		*Step A only*
			1	*2*	
Informative at both levels					
$n_j = 38$	0.437	0.496	0.506	0.503	0.503
$n_j = 0.4N_j$	0.420	0.494	0.510	0.503	0.503
$n_j = 9$	0.432	0.475	0.558	0.527	0.527
$n_j = 0.1N_j$	0.414	0.460	0.559	0.521	0.520
Informative only at level 2					
$n_j = 38$	0.500	0.493	0.499		0.499
$n_j = 0.4N_j$	0.500	0.491	0.501		0.501
$n_j = 9$	0.501	0.450	0.501		0.501
$n_j = 0.1N_j$	0.503	0.441	0.503		0.503
Non-informative					
$n_j = 38$	0.500	0.493	0.500		0.500
$n_j = 0.4N_j$	0.500	0.491	0.501		0.433
$n_j = 9$	0.501	0.451	0.500		0.500
$n_j = 0.1N_j$	0.500	0.438	0.499		0.424

†The true value of σ^2 is 0.5; the number of sampled level 2 units is $m = 35$; the number of replications is 1000.

biases of the unweighted estimators, as under scheme (a), this inflation is negligible compared with the corresponding decrease in bias. Note that for scheme (c), where weighting is redundant, the inflation in standard errors is the smallest. The standard errors of the three weighted estimators are generally very similar. The standard error estimators perform extremely well, with remarkably little bias except in the case of the standard error of the estimator $\hat{\sigma}^2$ obtained using step A only.

8. Application: survey of psychiatric morbidity

We now return to the example introduced in Section 1. We take the response variable to be the score on the clinical interview schedule—revised (CISR). This schedule is made up of 14 sections, each section covering a particular area of neurotic symptoms. 13 sections are scored with integer values from 0 to 4 and one section from 0 to 5. More frequent and more severe symptoms result in higher scores. The overall CISR value obtained by summing scores across the sections is a measure of psychiatric morbidity and takes integer values from 0 to 57. Values of 12 and above are taken to indicate significant psychiatric morbidity (Meltzer *et al.*, 1995).

We study the dependence of the CISR score on the following covariates, allowing for variation both within and between postal sectors:

age, 0 (under 40 years) or 1 (over 40 years); *sex*, 0 (female) or 1 (male); *work*, 0 (not working) or 1 (working); *housing tenure*, 0 (renter) or 1 (owner); *urban*, 0 (not urban) or 1 (urban); *qualifications*, 0 (A-level and above) or 1 (other).

In addition, we consider two size variables: S_j, *delivery point count* the number of delivery points in postal sector j; A_{ij}, the number of eligible adults at the delivery point containing person i in sector j.

Table 4. Simulation standard deviations of point estimators†

Sampling design	*Unweighted estimator*	*Weighted estimators*		
		Unscaled	*Scaled (method 2)*	*Step A only*
		Estimation of β		
Informative at both levels				
$n_j = 38$	78 (77)	90 (85)	89 (85)	89 (85)
$n_j = 0.4N_j$	76 (75)	90 (86)	90 (86)	90 (86)
Informative only at level 2				
$n_j = 38$	74 (75)	86 (85)	86 (85)	86 (85)
$n_j = 0.4N_j$	75 (76)	87 (86)	87 (86)	87 (86)
Non-informative				
$n_j = 38$	79 (77)	85 (82)	85 (82)	85 (82)
$n_j = 0.4N_j$	79 (78)	85 (84)	85 (83)	85 (83)
		Estimation of ω^2		
Informative at both levels				
$n_j = 38$	49 (47)	54 (49)	54 (49)	54 (52)
$n_j = 0.4N_j$	47 (45)	55 (50)	55 (50)	55 (53)
Informative only at level 2				
$n_j = 38$	50 (45)	58 (49)	58 (49)	58 (52)
$n_j = 0.4N_j$	51 (46)	58 (50)	58 (50)	58 (53)
Non-informative				
$n_j = 38$	50 (48)	52 (49)	52 (49)	52 (51)
$n_j = 0.4N_j$	51 (48)	55 (50)	54 (50)	55 (51)
		Estimation of σ^2		
Informative at both levels				
$n_j = 38$	19 (18)	26 (25)	24 (23)	24 (42)
$n_j = 0.4N_j$	19 (19)	27 (26)	27 (26)	30 (45)
Informative only at level 2				
$n_j = 38$	20 (20)	22 (22)	21 (21)	21 (40)
$n_j = 0.4N_j$	22 (21)	23 (22)	23 (23)	27 (43)
Non-informative				
$n_j = 38$	20 (20)	20 (19)	21 (21)	21 (40)
$n_j = 0.4N_j$	20 (20)	21 (21)	21 (21)	22 (32)

†Means of estimated standard errors are given in parentheses; all values are multiplied by 1000.

The dependence of CISR value on A_{ij} appears to be mainly according to whether there is one or more adults (the marginal CISR means are 7.0, 5.3, 5.2 and 5.5 for $A_{ij} = 1, 2, 3$ and 4 respectively) and so we define the additional variable *adults*, taking the values 0 ($A_{ij} \geqslant 2$) or 1 ($A_{ij} = 1$).

Initial attempts to fit the multilevel model (1) to these data resulted in residuals which were far from normal. We therefore applied the transformation $y = (\text{CISR score})^{1/2}$ which approximately produces normal residuals and removes the heteroscedasticity present when y is taken as the raw CISR score. We fit the following random intercept model to the transformed y-variable for various choices of covariate vectors x_{ij}:

$$y_{ij} = x_{ij}\beta + u_j + v_{ij}, \qquad u_j \sim N(0, \omega^2), \qquad v_{ij} \sim N(0, \sigma^2).$$

We computed the unweighted IGLS estimator and three PWIGLS estimators (unscaled, scaled by method 2 and scaled by method 2 with step A only). The variance estimators defined in Section 6 were used to provide standard errors. The weights w_j and $w_{i|j}$ were computed as described earlier from the π_j, which are proportional to the sizes S_j, and the $\pi_{i|j}$, which are products of the sample selection probabilities, $\pi^s_{i|j} = 90/S_j A_{ij}$, and the response probabilities $\pi^r_{i|j}$ given sample selection: $\pi_{i|j} = \pi^s_{i|j}\pi^r_{i|j}$. Among the 18000 selected delivery points, the number of responding adults was 10108. See Meltzer *et al.* (1995) for a description of the calculation of the response probabilities. In addition to unit non-response, there is also some item non-response and we only used data on the 9608 adults with complete responses. We scaled the response probabilities of Meltzer *et al.* (1995) accordingly so that $\Sigma^s_i w_{i|j}$ unbiasedly estimates N_j under the assumption of completely random item non-response within sectors and completely random unit non-response within the response weighting groups. We treat the resulting values of $\pi_{i|j}$ as given, ignoring possible error in the estimation of the $\pi^r_{i|j}$. The resulting weights w_j and $w_{i|j}$ have means 38.3 and 147.2 and standard deviations 20.2 and 93.2 respectively.

Our attempt at finding a parsimonious model led to the choice of covariates in the first model in Table 5. The model includes main effects for the two size variables and six other covariates together with three two-way interactions. Irrespective of the estimation method used, there is strong evidence of both significant covariate effects and significant between-area differences, as reflected by the estimators of ω^2. There is, however, no evidence of any effect of

Table 5. Estimates for the psychiatric morbidity data†

Parameter	*Unweighted estimator*	*Weighted estimators*		
		Unscaled	*Scaled (method 2)*	*Step A only*
Model including size variables				
β constant	2.30 (0.26)	2.27 (0.21)	2.23 (0.26)	2.23 (0.26)
Age	−0.12 (0.04)	−0.09 (0.04)	−0.10 (0.04)	−0.10 (0.04)
Sex	−0.28 (0.09)	−0.33 (0.08)	−0.29 (0.08)	−0.29 (0.08)
Work	−0.21 (0.05)	−0.24 (0.06)	−0.24 (0.06)	−0.24 (0.06)
Housing tenure	−0.24 (0.06)	−0.22 (0.04)	−0.20 (0.07)	−0.20 (0.07)
Urban	0.11 (0.06)	0.07 (0.06)	0.13 (0.06)	0.13 (0.06)
Qualifications	−0.02 (0.05)	−0.02 (0.03)	−0.03 (0.05)	−0.03 (0.05)
Adults	0.18 (0.07)	0.20 (0.06)	0.25 (0.08)	0.25 (0.08)
Delivery point count/1000	0.01 (0.04)	0.02 (0.06)	0.02 (0.05)	0.02 (0.05)
Work × sex	−0.24 (0.10)	−0.18 (0.09)	−0.22 (0.10)	−0.22 (0.10)
Adults × qualifications	0.17 (0.08)	0.16 (0.07)	0.16 (0.09)	0.16 (0.09)
Adults × age	−0.13 (0.07)	−0.15 (0.07)	−0.18 (0.09)	−0.18 (0.09)
ω^2	0.069 (0.017)	0.117 (0.017)	0.070 (0.016)	0.070 (0.017)
σ^2	2.053 (0.034)	1.956 (0.035)	1.999 (0.041)	1.959 (0.083)
Model excluding size variables				
β constant	2.41 (0.14)	2.39 (0.08)	2.34 (0.15)	2.34 (0.15)
Age	−0.14 (0.04)	−0.11 (0.04)	−0.11 (0.04)	−0.11 (0.04)
Sex	−0.30 (0.09)	−0.35 (0.08)	−0.31 (0.08)	−0.32 (0.08)
Work	−0.23 (0.05)	−0.25 (0.06)	−0.25 (0.06)	−0.25 (0.06)
Tenure	−0.28 (0.06)	−0.24 (0.04)	−0.23 (0.08)	−0.23 (0.08)
Urban	0.12 (0.09)	0.07 (0.06)	0.15 (0.09)	0.15 (0.09)
Qualifications	−0.03 (0.04)	−0.00 (0.04)	−0.01 (0.05)	−0.01 (0.05)
Work × sex	−0.22 (0.11)	−0.16 (0.09)	−0.20 (0.10)	−0.20 (0.10)
ω^2	0.071 (0.017)	0.118 (0.018)	0.071 (0.017)	0.071 (0.017)
σ^2	2.061 (0.033)	1.961 (0.035)	2.006 (0.042)	1.966 (0.084)

†Standard errors are given in parentheses.

the size variable S_j, the number of delivery points in the sector, with any of the four estimators. The apparent non-informativeness of the sampling of the postal sectors is further supported by the closeness of the unweighted and scaled weighted estimates of ω^2. A similar result was observed in the simulation study for the non-informative schemes at level 2 (see Table 2). In contrast, the effect of applying the unscaled weights is to increase the estimate of ω^2 considerably. Such increases were also observed in Table 2. It follows from equation (12) that when the N_j are large, as in our case, the effect of the second method of scaling may be to reduce $\hat{\omega}^2$ by roughly the average of $\hat{\sigma}^2/n_j$ across j (when $z_{ij} \equiv z_{0ij} \equiv 1$ as here). The sample mean of the $1/n_j$ is here 0.022 and $\hat{\sigma}^2 \doteq 2$ so the observed difference $0.117 - 0.070 = 0.047$ between the unscaled and scaled weighted estimators of ω^2 is indeed close to $2 \times 0.022 \doteq 0.044$. As in the simulation study, the weighted estimates employing step A only are very similar to the scaled weighted estimates, except for the estimate of σ^2 and its associated standard error. Our tentative interpretation is that the scaled weighted estimator is the least biased and thus preferable although more numerical evidence is desirable.

The effect of weighting on the estimated β-coefficients nowhere exceeds one standard error, but a large effect is not expected since we have included in our model the size variables S_j and A_{ij} which largely determine the selection probabilities. The four estimates for a given coefficient have always the same sign and are generally very similar. Note also that, unlike the results in Table 4, the standard errors of the weighted and unweighted estimators are very similar; this could result from a larger sample and smaller variation of the level 2 weights.

As noted in Section 1, we might expect the sample selection process not to lead to bias, if the model is well specified and includes as covariates the variables determining the sampling rates. To examine this further, we exclude the covariates which involve S_j and A_{ij}, to represent what could arise if these variables were unavailable or dropped from the model on substantive grounds. The results are given in the second part of Table 5.

We see again that (scaled) weighting has no effect on the estimate of ω^2, suggesting that the sampling of sectors is not informative with respect to y. The effect on the other parameter estimates is also not substantial, although there are reasons to believe that some of the differences represent selection effects rather than sampling error. For example, in the first model the presence of the adults × age interaction means that the (scaled) weighted estimated decrease in the mean of y for a person over 40 years of age is 0.10 if the person lives with other adults but 0.28 ($= 0.10 + 0.18$) if not. Similar decreases are estimated by the unweighted and unscaled estimators. The corresponding unweighted and scaled weighted estimates of the age coefficient in the model, excluding size variables 0.14 and 0.11 respectively, represent 'average effects' across the categories of the adults variable but in different proportions. The unweighted estimate attaches greater weight to one-adult households since these are oversampled. The scaled weighted estimate corrects for this disproportionate sampling and, since the age effect is lower for adults living with other adults, the weighted estimate of the age coefficient is lower than the unweighted estimate and in fact very close to the estimate in the first model for adults living with other adults.

9. Conclusions

Unequal probabilities of selection at any level of a hierarchical sampling scheme may bias standard estimators of parameters in an associated multilevel model. In particular, bias may even arise for standard 'self-weighting' designs where all level 1 units have equal overall inclusion probabilities, if higher level units have unequal selection probabilities. It is often possible to control for such bias by including relevant 'design variables' as covariates in the

multilevel model, but this may not be possible because of data availability or not be desirable for scientific reasons.

In this paper we consider two approaches to weighting IGLS estimators for multilevel models. The first approach uses reciprocals of selection probabilities and follows the broad principles of the pseudolikelihood approach. The second approach scales the weights in one of two ways. We also consider a simplified version of the second approach, implemented by applying the standard IGLS algorithm to a transformation of the data.

All three approaches are successful in removing the bias in the estimation of β. The first approach provides approximately unbiased and consistent estimators of the variance component parameters, but bias may arise when the level 1 sample sizes are small. Scaling helps to reduce this bias, especially when sampling is non-informative at level 1. When sampling is informative at level 1, scaling can overcorrect the bias although the second method of scaling generally seems preferable to no scaling. We have not identified any major effects of scaling on efficiency in our limited simulation study. Applying the standard IGLS algorithm after transforming the data is found to perform very similarly to scaled weighting in most cases, but when the level 1 sample sizes are related to the level 2 weights some bias seems to arise. It therefore seems difficult to recommend this as a general approach.

We tentatively recommend the weighted scaling method 2 as a means of reducing bias caused by informative sampling. In our simulation study these estimators perform fairly well and the associated variance estimators display remarkably little bias. We emphasize, however, that this study has only considered a limited set of possible forms of informative sampling and only a simple multilevel model. Even under these circumstances, some significant biases in the estimation of the level 2 variance arise when the level 1 sample sizes are small.

There appears to be little disadvantage in terms of bias or precision in using the scaled weighted estimators when sampling is non-informative. However, given the wide availability of unweighted estimators in standard multilevel modelling software, it will still be of interest in practice for survey data analysts to know whether there is a need for weighting, i.e. whether the sampling is informative. Some approaches to testing for informative sampling in single-level models are considered by Pfeffermann (1993) and Skinner (1994). These approaches might be extended to multilevel models. In this case it may be useful to test informativeness at each level and then to consider approaches which weight only at levels that are judged informative.

A final point relates to the performance of the variance estimators. As shown in Table 4, the estimators proposed perform very well for all the sampling schemes and estimators considered. The computation of these estimators is very simple even under complex sampling schemes. The use of the method 2 scaled estimators when the selection at both levels is with equal probabilities corresponds to the classical use of the standard IGLS estimators and it would be interesting to compare the performance of these estimators with the performance of variance estimators derived from the estimated information matrix.

Acknowledgements

This research was supported by the Economic and Social Research Council's Analysis of Large and Complex Datasets Programme. We thank the Office for National Statistics for making available the data from the survey of psychiatric morbidity. Special thanks are due to the referees for some very helpful comments.

Appendix A: Probability-weighted iterative generalized least squares estimation when $q > 1$

We indicate here how the theory of Section 3 extends to the case $q > 1$. Replacing a_j (defined below equation (6)) by $A_j = (Z_j'D_j^{-1}Z_j + \hat{\sigma}^2\hat{\Omega}^{-1})^{-1}$, where $\hat{\sigma}^2$ and $\hat{\Omega}$ are the IGLS census estimates from iteration $r-1$, we note that V_{jr} defined below equation (3) satisfies

$$V_{jr}^{-1} = \hat{\sigma}^{-2}D_j^{-1} - \hat{\sigma}^{-2}D_j^{-1}Z_jA_jZ_j'D_j^{-1}. \tag{16}$$

Hence, the terms in equation (3) can be expressed as

$$P^{(r)} = \sum_j (X_j'D_j^{-1}X_j - X_j'D_j^{-1}Z_jA_jZ_j'D_j^{-1}X_j),$$

$$Q^{(r)} = \sum_j (X_j'D_j^{-1}Y_j - X_j'D_j^{-1}Z_jA_jZ_j'D_j^{-1}Y_j).$$

Given $\hat{\sigma}^2$ and $\hat{\Omega}$, stage 1 of the IGLS algorithm depends therefore only on the 'sufficient statistics' $X_j'D_j^{-1}X_j$, $X_j'D_j^{-1}Y_j$, $X_j'D_j^{-1}Z_j'D_j^{-1}Y_j$ and $Z_j'D_j^{-1}Z_j$ $(j = 1, \ldots, M)$. As for the case $q = 1$, each of these terms may be expressed as a sum over i, e.g. $X_j'D_j^{-1}X_j = \Sigma_i\, x_{ij}x_{ij}'/z_{0ij}^2$. Turning to stage 2, note first from equation (16) and the definition of A_j that $V_{jr}^{-1}G_{kj} = \hat{\sigma}^{-2}\delta_{ks}I_{N_j} + \hat{\sigma}^{-2}D_j^{-1}Z_jB_{kj}Z_j'$, where G_{kj} is defined below equation (2), I_{N_j} is the $N_j \times N_j$ identity matrix and $B_{kj} = \hat{\sigma}^2A_j\hat{\Omega}^{-1}H_{kj} - \delta_{ks}A_j$. Letting $C_{kj} = -\delta_{ks}A_j + B_{kj} - B_{kj}Z_j'D_j^{-1}Z_jA_j$, the klth element of $R^{(r)}$ in equation (4) can be expressed as

$$\sum_j \{\delta_{ks}\delta_{ls}N_j + \delta_{ls}\,\mathrm{tr}(Z_j'D_j^{-1}Z_jC_{kj}) + \delta_{ks}\,\mathrm{tr}(Z_j'D_j^{-1}Z_jH_{lj}) + \mathrm{tr}(Z_j'D_j^{-1}Z_jC_{kj}Z_j'D_j^{-1}Z_jH_{lj})\} \tag{17}$$

and the kth element of $S^{(r)}$ can be expressed as

$$\sum_j [\delta_{ks}\,\mathrm{tr}\{(Y_j - X_j\beta)'D_j^{-1}(Y_j - X_j\beta)\} + \mathrm{tr}\{(Y_j - X_j\beta)'D_j^{-1}Z_jC_{kj}Z_j'D_j^{-1}(Y_j - X_j\beta)\}].$$

It follows that stage 2 depends on the same sufficient statistics as stage 1 and also on the sizes N_j of the level 2 units. PWIGLS estimation may again be achieved by applying step A to the sample data and modifying the resulting IGLS algorithm by step B which now becomes

(a) insert w_j into the sample sum corresponding to equation (17) and
(b) replace n_j in the first term in the sample version of equation (17) (this term is $\delta_{ks}\delta_{ls}n_j$) by $\hat{N}_j = \Sigma_i^s\, w_{i|j}$.

References

Anderson, T. W. (1973) Asymptotically efficient estimation of covariance structures with linear structure. *Ann. Statist.*, **1**, 135–141.

Binder, D. A. (1983) On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.*, **51**, 279–292.

Duncan, C., Jones, K. and Moon, G. (1995) Psychiatric morbidity: a multilevel approach to regional variations in the UK. *J. Epidem. Commty Hlth*, **49**, 290–295.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43–56.

——— (1995) *Multilevel Statistical Models*, 2nd edn. London: Arnold.

Graubard, B. I. and Korn, E. L. (1996) Modelling the sampling design in the analysis of health surveys. *Statist. Meth. Med. Res.*, **5**, 263–281.

Isaki, C. T. and Fuller, W. A. (1982) Survey design under the regression super-population model. *J. Am. Statist. Ass.*, **77**, 89–96.

Longford, N. T. (1995) Model-based methods for analysis of data from 1990 NAEP Trial State Assessment. *Report NCES 95-696*. National Center for Education Statistics, Washington DC.

Meltzer, H., Gill, B., Petticrew, M. and Hinds, K. (1995) *The Prevalence of Psychiatric Morbidity among Adults Living in Private Households*. London: Her Majesty's Stationery Office.

Pfeffermann, D. (1993) The role of sampling weights when modelling survey data. *Int. Statist. Rev.*, **61**, 317–337.

Pfeffermann, D. and LaVange, L. M. (1989) Regression models for stratified multi-stage cluster samples. In *Analysis of Complex Surveys* (eds C. J. Skinner, D. Holt and T. M. F. Smith), pp. 237–260. Chichester: Wiley.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Shah, B. V. and LaVange, L. M. (1994) Mixed models for survey data. *Joint Statistical Meet., Aug.*
Skinner, C. J. (1989) Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (eds C. J. Skinner, D. Holt and T. M. F. Smith), pp. 59–87. Chichester: Wiley.
———(1994) Sample models and weights. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 133–142.