Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2002/01

Weighted estimation in multilevel ordinal models to allow for informativeness of the sampling design

Leonardo Grilli, Monica Pratesi

Università degli Studi di Firenze

Statistics

Weighted estimation in multilevel ordinal models to allow for informativeness of the sampling design

Leonardo Grilli
Dipartimento di Statistica
Università di Firenze
grilli@ds.unifi.it

Monica Pratesi
Dipartimento di Matematica, Statistica, Informatica e Applicazioni
Università di Bergamo
pratesi@unibg.it

January 15, 2002

Abstract

Multilevel ordinal models are often fitted to survey data gathered with a complex multistage sampling design. However, if such a design is informative, in the sense that the inclusion probabilities depend on the response variable even after conditioning on the covariates, then standard maximum likelihood estimators are biased. In this paper, following the Pseudo Maximum Likelihood (PML) approach of Skinner (1989), we propose a probability-weighted estimation procedure for multilevel ordinal models which eliminates the bias generated by the informativeness of the design. The reciprocals of the inclusion probabilities at each sampling stage are used to weight the log-likelihood function and the weighted estimators obtained in this way are tested by means of a simulation study. The variance estimators are obtained by a bootstrap procedure. The maximization of the weighted log-likelihood of the model is done by the NLMIXED procedure of the SAS, which is based on adaptive Gaussian quadrature. Also the bootastrap estimation of variances is implemented in the SAS environment.

Keywords: Multilevel ordinal model, Multistage sampling, Pseudo Maximum Likelihood.

1 Introduction

Multilevel models for ordinal responses, including binary responses as a special case, are frequently used in many areas of research for modelling hierarchically clustered populations. In fact, both in human and biological sciences, the status or the response of a subject may often be classified in two categories or in a set of ordered categories (ordinal or graded scale). At the same time, subjects are observed clustered in groups (e.g. schools, firms, clinics, geographical areas). The hierarchical population structure is often also employed to design multi-stage sampling schemes, with unequal selection probabilities at some or all the stages of the sampling process. In the multilevel analysis of survey data, complex sampling schemes are often ignored even if they may cause the violation of the basic assumptions underlying multilevel models. In fact, in complex sampling designs both the subjects and the clusters at all levels could be selected with probabilities that, even conditionally on the covariates, do depend on the response variable; in other words, the sampling design might be informative.

For data that are clustered and obtained by multistage informative designs, proposals for fitting multilevel models have been formulated only for the case of continuous response variable. In a recent article, Pfeffermann *et al.* (1998) propose probability-weighting procedures of first and second level units that adjust for the effect of an informative design on the estimation of two-level models with continuous response variable. The method, known as Pseudo Maximum Likelihood (PML), consists in writing down a closed form expression for the census likelihood, estimating the log-likelihood function and then maximizing the estimated function numerically. The method needs the sampling weights for the sampled elements and clusters at all levels. The authors also develop appropriate 'sandwich' estimates for the variances of the estimators.

The wide use of multilevel ordinal and binary models in many fields of application urges for an analogous solution, which should be both effective and simple to implement, preferably in the framework of a standard statistical software. The present paper represent a contribution in this direction.

The paper structure is as follows. Basic definitions for the multilevel ordinal model are set out in Section 2, while in Section 3 the probability-weighting approach for fitting the model is developed. In section 4 the properties of the various estimators are evaluated by a simulation study. Section 5 contains some final remarks.

2 The multilevel ordinal model

In order to ease the comparison with the results concerning the linear model (Pfeffermann *et al.*, 1998), it is useful to write the ordinal model in terms of a latent linear model endowed with a set of thresholds. Suppose that an observed ordinal response variable Y, with $k = 1, 2, \ldots, K$ levels, is generated, through a set of thresholds, by a latent continuous variable $\tilde{Y}$ following a variance component model (Hedeker and Gibbons, 1994):

$$\tilde{Y}_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \omega u_j + \varepsilon_{ij}, \tag{1}$$

with $i = 1, 2, \ldots, N_j$ elementary units for the j-th cluster ($j = 1, 2, \ldots, M$). In (1) $\mathbf{x}_{ij}$ is a covariate vector and $\boldsymbol{\beta}$ is the corresponding vector of slopes; the random variables ε_{ij} and u_j are the disturbances, respectively at the first (elementary) and second (cluster) level; and ω^2 is the second level variance component.

For the disturbances of model (1) we make the standard assumptions, i.e. a) the ε_{ij}'s are iid with zero mean and unknown variance σ^2; b) the u_j's are Gaussian iid with zero mean and unit variance; c) the ε_{ij}'s and u_j's are mutually independent.

Note that model (1) leads to the simplest case of multilevel ordinal model, with just two levels and a single random effect on the intercept; the extension to three or more levels and to multiple random effects is straightforward in principle (Gibbons and Hedeker, 1997), but the complications in the formulae suggest to consider only the simplest case, which is sufficient for the discussion of the main conceptual issues.

The observed ordinal variable Y is linked to the latent one $\tilde{Y}$ through the following relationship:

$$\{Y_{ij} = k\} \quad \Leftrightarrow \quad \left\{\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k\right\},$$

where the thresholds satisfy $-\infty = \gamma_0 \leq \gamma_1 \leq \ldots \leq \gamma_{K-1} \leq \gamma_K = +\infty$. Therefore, conditional on u_j, the model probability for subject i of cluster j is

$$\begin{aligned} P\left(Y_{ij} = k \,|\, u_j\right) &= P\left(\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k \,|\, u_j\right) \\ &= P\left(\tilde{Y}_{ij} \leq \gamma_k \,|\, u_j\right) - P\left(\tilde{Y}_{ij} \leq \gamma_{k-1} \,|\, u_j\right), \end{aligned} \tag{2}$$

with

$$\begin{aligned} P\left(\tilde{Y}_{ij} \leq \gamma_k \,|\, u_j\right) &= P\left(\varepsilon_{ij} \leq \gamma_k - \left[\boldsymbol{\beta}'\mathbf{x}_{ij} + \omega u_j\right] \,|\, u_j\right) \\ &= F\left(\frac{\gamma_k}{\sigma} - \left[\frac{1}{\sigma}\boldsymbol{\beta}'\mathbf{x}_{ij} + \frac{\omega}{\sigma}\, u_j\right]\right) \\ &= F\left(\gamma_{\sigma,k} - \left[\boldsymbol{\beta}'_\sigma\mathbf{x}_{ij} + \omega_\sigma\, u_j\right]\right), \end{aligned} \tag{3}$$

where $F(\cdot)$ is the distribution function of the standardized first level error term ε_{ij}/σ. All the model parameters are defined in terms of the unknown σ, the standard deviation of the first level error term, so only the ratios of the model parameters to the standard deviation of the first level error term are identifiable; we use the notation ψ_σ to indicate that the latent model parameter ψ is in σ units, i.e. $\psi_\sigma \equiv \psi/\sigma$. Note that $F(\cdot)$ is also the inverse of the link function of the ordinal model: for example, the standard Gaussian distribution function yields the ordinal probit model.

Now let $\boldsymbol{\theta}$ denote the vector of all estimable parameters, which include $\boldsymbol{\beta}_\sigma$, ω_σ and $K-2$ thresholds $\{\gamma_{\sigma,k} : k = 2, \ldots, K-1\}$ ($\gamma_{\sigma,1}$ is fixed to zero to insure identifiability). The conditional likelihood for subject i of cluster j is

$$L_{ij}(\boldsymbol{\theta} \mid u) = \prod_{k=1}^{K} \left[P\left(Y_{ij} = k \mid u_j\right)\right]^{d_{ijk}}, \tag{4}$$

where $P\left(Y_{ij} = k \mid u_j\right)$ is defined by (2) and (3), while d_{ijk} is the indicator function of the event $\{Y_{ij} = k\}$. Then the marginal likelihood for cluster j is

$$L_j(\boldsymbol{\theta}) = \int_{-\infty}^{+\infty} \prod_{i=1}^{N_j} L_{ij}(\boldsymbol{\theta} \mid u)\phi(u)du,$$

where ϕ is the standard Gaussian density function. Finally, the overall marginal likelihood is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{M} L_j(\boldsymbol{\theta}). \tag{5}$$

3 Probability-weighted estimation

3.1 Pseudo Maximum Likelihood (PML) estimators

Suppose that the whole population of M clusters (level 2 units) with N_j elementary units (level 1 units) per cluster is not observed; instead the following two-stage sampling scheme is used:

- first stage: m clusters are selected with inclusion probabilities π_j ($j = 1, \ldots, M$);
- second stage: n_j elementary units are selected within the j-th selected cluster with probabilities $\pi_{i|j}$ ($i = 1, \ldots, N_j$).

The unconditional sample inclusion probabilities are then $\pi_{ij} = \pi_{i|j}\pi_j$.

When the sampling mechanism is informative, i.e. the π_j and/or the $\pi_{i|j}$ depend on the model disturbances and hence on the response variable, the maximum likelihood estimator of the parameters of the ordinal variance component model defined in Section 2 may be seriously biased.

A standard solution to this problem is the Pseudo Maximum Likelihood (PML) approach (Skinner, 1989). However in the context of multilevel models the implementation of the PML approach is complicated by the fact that the population log-likelihood is not a simple sum of elementary unit contributions, but rather a function of sums across level 2 and level 1 units. This can be seen by writing the logarithm of the likelihood (5) as follows:

$$\log L(\boldsymbol{\theta}) = \sum_{j=1}^{M} \log \int_{-\infty}^{+\infty} \left[\exp\left\{ \sum_{i=1}^{N_j} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \right] \phi(u)du. \tag{6}$$

A design-consistent estimate of the population log-likelihood (6) can be obtained applying the Horvitz-Thompson principle, i.e. replacing each sum over the level 2 population units j by a sample sum weighted by $w_j \equiv 1/\pi_j$ and each sum over the level 1 units i by a sample sum weighted by $w_{i|j} \equiv 1/\pi_{i|j}$:

$$\log \hat{L}(\boldsymbol{\theta}) = \sum\nolimits_j^s w_j \log \int_{-\infty}^{+\infty} \left[\exp \left\{ \sum\nolimits_i^s w_{i|j} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \right] \phi(u) du, \qquad (7)$$

where $\sum^s$ denotes a sum over sample units.

Note that inserting the weights in the log-likelihood implies the use of a design-consistent estimator of the population score function. In fact, the population score function $U(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})$ can be written as

$$\sum_{j=1}^{M} \frac{\int_{-\infty}^{+\infty} \left[\exp \left\{ \sum_{i=1}^{N_j} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \right] \cdot \left\{ \sum_{i=1}^{N_j} \frac{\partial}{\partial \boldsymbol{\theta}} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \phi(u) du}{\int_{-\infty}^{+\infty} \exp \left\{ \sum_{i=1}^{N_j} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \phi(u) du}, \qquad (8)$$

whose corresponding Horvitz-Thompson estimator $\hat{U}(\boldsymbol{\theta})$ is

$$\sum\nolimits_j^s w_j \frac{\int_{-\infty}^{+\infty} \left[\exp \left\{ \sum_i^s w_{i|j} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \right] \cdot \left\{ \sum_i^s w_{i|j} \frac{\partial}{\partial \boldsymbol{\theta}} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \phi(u) du}{\int_{-\infty}^{+\infty} \exp \left\{ \sum_i^s w_{i|j} \log L_{ij}(\boldsymbol{\theta} \mid u) \right\} \phi(u) du}, \qquad (9)$$

which equals the score obtained by differentiating the probability-weighted log-likelihood (7).

Under mild conditions, the solution $\hat{\boldsymbol{\theta}}_{PML}$ to the estimating equation $\hat{U}(\boldsymbol{\theta}) = \mathbf{0}$ is design-consistent for the finite-population maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ which, in turn, is model-consistent for the super-population parameter $\boldsymbol{\theta}$: therefore $\hat{\boldsymbol{\theta}}_{PML}$ is a consistent estimator of $\boldsymbol{\theta}$ with respect to the mixed design-model distribution (Pfeffermann, 1993).

The implementation of the PML approach requires the knowledge of the inclusion probabilities at both levels. Using only second level weights or only first level weights may be insufficient or may even worsen the situation, as shown by our simulations.

3.2 Scaling the weights

A controversial issue present in Pfeffermann *et al.* (1998) is the scaling of the weights to obtain estimators with little bias even in small samples. Obviously, scaling is not relevant for the level 2 weights, since from (7) and (9) it is clear that multiplying the w_j's by a constant does not change the PML estimates (it simply inflates the information matrix by that constant). On the contrary, scaling the level 1 weights may have important effects on the small sample behavior of the PML estimator. In the simulation study discussed in Section (4) we present the results for the following type of scaling (named 'scaling method

2' in Pfeffermann *et al.*, 1998):

$$w_{i|j}^{scaled} = \frac{w_{i|j}}{\bar{w}_j}, \tag{10}$$

where $\bar{w}_j = (\sum_i^s w_{i|j})/n_j$, so that, for the j-th cluster, the sum of the scaled weights equals the cluster sample size n_j. In the present paper we do not wish to discuss the relative merits of the various scaling methods, so we limit our simulations to scaled weights (10), which have an intuitive meaning and showed a good performance in the study of Pfeffermann *et al.* (1998).

3.3 Estimation technique

The maximization of the weighted log-likelihood (7) involves the computation of several integrals which do not have a closed-form solution, so a numerical approximation technique is required. When the dimensionality of the integrals is low, a simple and very accurate technique is Gaussian quadrature, which is based on a summation over an appropriate set of points. The NLMIXED procedure of the SAS (SAS Institute, 1999) is a general procedure for fitting nonlinear random effects models using adaptive Gaussian quadrature. Various optimization techniques are available to carry out the maximization; the default, used in the simulations of Section 4, is a dual quasi-Newton algorithm, where dual means that the upgrading concerns the Cholesky factor of an approximate Hessian (SAS Institute, 1999).

Though the NLMIXED procedure does not include an option for PML estimation, it is still possible to insert the weights in the likelihood, using different tricks for level 1 and level 2 weights. To insert level 1 weights it is necessary to exploit the option which allows to write, with SAS programming statements, an arbitrary expression for the conditional likelihood of the model: then one should simply replace $L_{ij}(\boldsymbol{\theta} \mid u)$ of formula (4) with $w_{i|j} L_{ij}(\boldsymbol{\theta} \mid u)$. On the other hand, level 2 weights can be inserted in the likelihood through the REPLICATE statement. Unfortunately, this statement is limited to integer weights, so to avoid gross approximations it is advisable to proceed as follows: a) inflate all the level 2 weights by an arbitrary constant k; b) insert the integer part of the inflated weights in the likelihood through the REPLICATE statement; c) rescale the information matrix by the constant k or, equivalently, multiply the estimation covariance matrix by k. This procedure relies on the fact that multiplying the level 2 weights by a constant has the only effect of inflating the information matrix by that constant, leaving the estimates unchanged.

3.4 Variance estimation

In standard maximum likelihood the estimation of the covariance matrix of the estimators is obtained by inverting the information matrix. However this conventional estimator is not appropriate when using the PML method. In this case Skinner (1989) proposed the use of a robust 'sandwich' estimator, which is employed also by Pfeffermann *et al.* (1998).

As described in Section 3.3, the NLMIXED procedure of the SAS allows to fit the model with the PML approach, but the estimated covariance matrix, which is obtained simply inverting the information matrix, is likely to be misleading in order to appreciate the actual variability of PML estimators. In the SAS framework a simple and effective solution, requiring a bit programming, is to empirically estimate the variance through the bootstrap technique for finite populations (Särndal *et al.*, 1992), which consists of the following steps: *a*) using the sample data, an artificial finite population is constructed, assumed to mimic the real population; *b*) a series of independent bootstrap samples is drawn from the artificial finite population and for each bootstrap sample an estimate of the target parameter is calculated; *c*) the bootstrap variance estimate is obtained as the variance of the observed distribution of the bootstrap estimates.

The artificial finite population can be generated in the following way: *i*) for the j-th sampled cluster, each of the n_j sampled elementary units is replicated $w_{i|j}$ times, rounding the weight to the nearest integer, obtaining an artificial cluster of about N_j elementary units; *ii*) each of the m artificial clusters is replicated w_j times, rounding the weight to the nearest integer, obtaining an artificial population of about M clusters. Then the samples are selected from the artificial population in the following way: *i*) m clusters are re-sampled with probability proportional to π_j; *ii*) for the j-th re-sampled cluster, n_j elementary units are re-sampled with probability proportional to $\pi_{i|j}$.

When the sampling fraction m/M is low, most of the variance is due to the sampling of the clusters, so the bootstrap procedure described above could be simplyfied by omitting the steps concerning the elementary units, i.e. step *i)* in the construction of the artificial population and step *ii)* in the re-sampling process.

4 Simulation study

4.1 Design of experiment

The experiment reflects the two-stage scheme assumed for the observed variables: first, the finite-population values are generated from the adequate superpopulation model (stage I) and then an informative or non-informative sample is selected from the finite population (stage II), with one sample per population. The two-stage selection scheme was repeated 1000 times for each combination of sample size and type of informativeness. In order to compare our results with the ones obtained for the multilevel linear model, the experiment has been designed following the example of Pfeffermann *et al.* (1998, section 7).

The simulation study was limited to the simplest case of the model defined in Section 2, with only two categories for the response variable (i.e. $K = 2$) and no explanatory variables; moreover, for the first-level disturbances a Gaussian distribution was assumed, so the fitted model is in fact a random intercept probit binary model.

The values of the binary response variable Y_{ij} were generated using the mentioned two-stage scheme:

- Stage I. Finite-population values Y_{ij} ($j = 1, \ldots, M$; $i = 1, \ldots, N_j$) were obtained by first generating a value from the super-population latent model $\tilde{Y}_{ij} = \beta + u_j + v_{ij}$, with $u_j \sim N\left(0, \omega^2\right)$ and $v_{ij} \sim N\left(0, \sigma^2\right)$, and then putting $Y_{ij} = 0$ if $\tilde{Y}_{ij} \leq 0$ or $Y_{ij} = 1$ if $\tilde{Y}_{ij} > 0$ (recall that the binary model has only one threshold which is set to zero to guarantee identifiability). The latent model parameter values employed in the simulation are $\beta = 0$, $\omega^2 = 0.2$ and $\sigma^2 = 0.5$, so that the parameters estimable from the binary model are $\beta_\sigma \equiv \beta/\sigma = 0$ and $\omega_\sigma \equiv \omega/\sigma = 0.632$ (see expression (3)). The hierarchical structure of the population comprises $M = 300$ clusters, while the cluster sizes N_j were determined, as in Pfeffermann *et al.* (1998), by $N_j = 75 \exp(\tilde{u}_j)$, with $\tilde{u}_j$ generated from $N\left(0, \omega^2\right)$, truncated below by -1.5ω and above by 1.5ω. As a result, in our population N_j lies in the range $[38, 147]$ with mean around 80.

- Stage II. Once the finite-population values were obtained, we adopted one of the following sampling schemes, as in Pfeffermann *et al.* (1998):

  (a) *Informative at both levels*: first, m clusters were selected with probability proportional to a 'measure of size' X_j, i.e. $\pi_j = mX_j / \sum_{j=1}^{M} X_j$; 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 elementary units in the j-th sampled cluster were then partitioned into two strata according to whether $v_{ij} > 0$ or $v_{ij} \leq 0$ 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 the previous one, except that simple random sampling was employed for the selection of level 1 units within each sampled cluster.

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

Following Pfeffermann *et al.* (1998), the simulation study included samples with $m = 35$ clusters and varying numbers of elementary units: large samples with fixed size $n_j = n_0 = 38$ and proportional allocation $n_j = 0.4N_j$, and small samples with fixed size $n_j = n_0 = 9$ and proportional allocation $n_j = 0.1N_j$ (mean of about 9).

The simulation study was carried out entirely within the SAS System (SAS Institute, 1999), writing a specific code with the macro language. The models were fitted with the NLMIXED procedure, using 10-point adaptive Gaussian quadrature with a dual quasi-Newton algorithm, which reached convergence in a few iterations. As explained in Section 3.3, to avoid gross rounding errors the level 2 weights were pre-multiplied by a factor $k = 10000$ and the estimation covariance matrix was then multiplied by the same factor.

4.2 Results

The results of the simulations are shown in Tables 1 and 2. For each sampling design the behavior of the point estimators of the intercept β_σ and the second level standard deviation ω_σ is summarized by the mean and standard deviation of their Monte Carlo sampling distribution. The point estimators under study are the standard maximum likelihood unweighted estimator and the following three weighted versions of it:

- only level 2 weights (i.e. varying w_j's and constant $w_{i|j}$'s);
- weights at both levels, with unscaled level 1 weights;
- weights at both levels, with level 1 weights scaled according to (10).

Our results are shown and discussed according to the following three sc enarios:

1. Base scenario: the sampling design is non-informative. In this situation all the basic assumptions underlying the random intercept binary model are fulfilled, so this case can be assumed as a benchmark for judging the subsequent results.
2. Unweighted scenario: the sampling design is informative and the estimators are unweighted. In this situation the basic assumptions underlying the random intercept binary model are violated because of the informativeness of the design and no adjustment is used.
3. Weighted scenario: the sampling design is informative and the estimators are weighted. Also in this case the basic assumptions underlying the random intercept binary model are violated, but the weights are introduced as a tentative adjustment for the bias of the estimators.

4.2.1 Base scenario

When the sampling design is non-informative the standard maximum likelihood unweighted estimator is asymptotically unbiased (Tables 1 and 2: row 3, column 1). However for small samples ($n_j = 9$ and $n_j = 0.1N_j$) there is an appreciable negative bias in the estimation of ω_σ.

If the weights are introduced when there is no need to adjust for the effect of the design (Tables 1 and 2: row 3, columns 2-4), we face with a slight increase in the variability of the estimators, which is more pronounced when the unscaled estimators are used with small samples. In that situation there is also a surprising behavior of the estimator of ω_σ, for which the sign of the bias is reversed.

Table 1: Simulation means and standard deviations (in parenthesis) of point estimators of the intercept (true value 0, number of replicates 1000)

Sampling design	*Unweighted estimator*	*Weighted estimators*		
		Only level 2 weights	*Weights at both levels - Unscaled*	*Weights at both levels - Scaled*
Informative at both levels				
Fixed size $n_j = 38$	-0.120 (0.212)	-0.411 (0.202)	0.014 (0.193)	0.015 (0.188)
Prop. size $n_j = 0.4N_j$	-0.163 (0.212)	-0.453 (0.200)	0.018 (0.190)	0.021 (0.183)
Fixed size $n_j = 9$	-0.214 (0.204)	-0.512 (0.190)	-0.062 (0.258)	0.000 (0.185)
Prop. size $n_j = 0.1N_j$	-0.164 (0.220)	-0.450 (0.209)	-0.074 (0.294)	0.008 (0.203)
Informative only at level 2				
Fixed size $n_j = 38$	0.281 (0.169)	0.018 (0.168)	0.017 (0.170)	0.017 (0.169)
Prop. size $n_j = 0.4N_j$	0.274 (0.169)	0.014 (0.178)	0.014 (0.182)	0.014 (0.181)
Fixed size $n_j = 9$	0.274 (0.187)	0.010 (0.195)	0.010 (0.212)	0.009 (0.196)
Prop. size $n_j = 0.1N_j$	0.269 (0.179)	0.007 (0.179)	0.007 (0.203)	0.006 (0.182)
Non-informative				
Fixed size $n_j = 38$	0.000 (0.108)	0.000 (0.114)	0.001 (0.115)	0.001 (0.115)
Prop. size $n_j = 0.4N_j$	0.003 (0.113)	0.004 (0.120)	0.003 (0.123)	0.003 (0.122)
Fixed size $n_j = 9$	-0.007 (0.108)	-0.009 (0.115)	-0.010 (0.125)	-0.010 (0.117)
Prop. size $n_j = 0.1N_j$	-0.002 (0.110)	-0.002 (0.114)	-0.004 (0.132)	-0.003 (0.117)

Table 2: Simulation means and standard deviations (in parenthesis) of point estimators of the second level standard deviation (true value 0.632, number of replicates 1000)

Sampling design	*Unweighted estimator*	*Weighted estimators*		
		Only level 2 weights	*Weights at both levels - Unscaled*	*Weights at both levels - Scaled*
Informative at both levels				
Fixed size $n_j = 38$	0.671 (0.106)	0.638 (0.112)	0.637 (0.137)	0.604 (0.128)
Prop. size $n_j = 0.4N_j$	0.673 (0.108)	0.636 (0.112)	0.645 (0.142)	0.592 (0.130)
Fixed size $n_j = 9$	0.644 (0.145)	0.584 (0.172)	0.920 (0.289)	0.536 (0.222)
Prop. size $n_j = 0.1N_j$	0.598 (0.164)	0.546 (0.183)	1.002 (0.317)	0.498 (0.242)
Informative only at level 2				
Fixed size $n_j = 38$	0.595 (0.100)	0.596 (0.110)	0.605 (0.111)	0.601 (0.111)
Prop. size $n_j = 0.4N_j$	0.582 (0.096)	0.582 (0.115)	0.603 (0.113)	0.596 (0.113)
Fixed size $n_j = 9$	0.547 (0.121)	0.548 (0.135)	0.671 (0.144)	0.563 (0.133)
Prop. size $n_j = 0.1N_j$	0.538 (0.122)	0.535 (0.142)	0.696 (0.158)	0.551 (0.139)
Non-informative				
Fixed size $n_j = 38$	0.611 (0.086)	0.612 (0.092)	0.621 (0.090)	0.617 (0.091)
Prop. size $n_j = 0.4N_j$	0.609 (0.084)	0.606 (0.088)	0.626 (0.088)	0.618 (0.088)
Fixed size $n_j = 9$	0.561 (0.105)	0.561 (0.112)	0.685 (0.119)	0.575 (0.111)
Prop. size $n_j = 0.1N_j$	0.551 (0.109)	0.546 (0.113)	0.703 (0.134)	0.559 (0.112)

4.2.2 Unweighted scenario

The informativeness of the sampling design produces biased and unstable estimates. The bias is still evident for large samples (Tables 1 and 2: rows 1-2, column 1). The conclusions are the same for both types of informative designs, though the bias has a different sign. Moreover the informativeness of the design inflates the variability of the estimators with respect to the base scenario: in particular, when the design is informative at both levels the standard error of the estimator of β_σ is doubled.

4.2.3 Weighted scenario

Estimation of β_σ.

The results in Table 1 show that, when the design is informative, the weighted based adjustment is effective in removing the bias in the estimation of β_σ.

Particularly, when the design is informative only at level 2 (Table 1: row 2, columns 2-4) and the weights are introduced only at this level, the bias in the estimation is corrected with no important increase in the sampling variance. The result is valid also for fully weighted estimators (unscaled or scaled). The bias correction works for small samples too.

When the design is informative at both levels (Table 1: row 1, columns 2-4) and the weights are introduced for both level 1 and level 2 units, the bias in the estimation of β_σ is corrected, also reducing the sampling variance. In small samples scaling is preferable, since it allows to achieve an unbiased estimator with a substantial lower sampling variance. It should be noted that when the design is informative at both levels, the estimator which uses only the level 2 weights is worse than the standard unweighted estimator.

Estimation of ω_σ.

The results in Table 2, concerning ω_σ, are more difficult to be interpreted (Table 2: rows 1-2, columns 2-4) . First note that also in the base scenario the estimation of ω_σ is biased, especially for small samples. Therefore the weight-based adjustment should be judged as effective if it is able to reproduce the same bias which is observed in the base scenario. On these grounds the behavior of the scaled weighted estimator is satisfactory in nearly all situations, with the exception of the small samples when the design is informative at both levels. In that case there is also a not negligible number of replications which yielded a zero estimate for ω_σ (4.5% for the design with fixed size and 2% for the design with proportional size). The unscaled weighted estimator do not suffer from the problem of null estimates, but, apart from having a larger variance than the scaled version, tends to overestimate ω_σ, showing a relative bias of about 50% in the small samples when the design is informative at both levels. Note also that the scaled estimator outperforms the estimator which uses only level 2 weights even when the design is informative only at level 2.

4.2.4 General remarks

Our simulations showed that the PML approach is, in most cases, a simple and effective strategy to deal with informative sampling designs. The only requirement is the knowledge of the inclusion probabilities at every stage of the sampling process (except when the informativeness does not concerns all the levels).

The scaled version of the weighted estimator can be recommended in general, since it allows to obtain a low bias with a modest increase in the sampling variance. Even if weighting is superfluous, the loss of efficiency due to the inclusion of scaled weights is very low. Moreover, if the design is informative at both levels, when passing from standard estimation to scaled weighted estimation the sampling variance increases for ω_σ, but even decreases for β_σ.

While for the estimation of β_σ weighting is always effective, for ω_σ attention should be paid to the sample size: in fact, weighting leads to satisfactory results only when the cluster size is high, i.e. when it allows a good representation of the complex variance structure. However the sample size is crucial in the estimation of ω_σ also when all the basic assumptions for the estimation of the multilevel ordinal model are satisfied.

The differences induced by the type of clusters in the sample, fixed or variable size, are minimal, with equal sized clusters leading to slightly better estimators; however, as already noted, the important differences are always due to the average size of the clusters in the sample.

The results of our simulation study confirm the findings of Pfeffermann *et al.* (1998) on the random intercept linear model: probability-weighted estimators are good for the intercept, while some relevant bias remains in the estimation of the variance component when the sample is small. As was to be expected, when passing from a linear to a non-linear model the performance of the estimators slightly worsen, but the direction and importance of the bias in the various cases are similar. Also the advantages of scaling are confirmed.

The critical point in the random intercept binary model is the estimation of the variance component ω_σ, which represents a difficult task also when the design is non-informative. Note that, using the threshold formulation, ω_σ is defined as ω/σ, so estimation of ω_σ conveys the problems that in the linear model are associated with the estimation of the two variance components.

Finally the comparison between our results and those of Pfeffermann *et al.* (1998) shows an interesting difference concerning the sampling variance of the weighted estimators. In particular, in the random intercept linear model the inclusion of the weights increases the sampling variance in a significant manner, while scaling has little role in reducing the variance. On the contrary, in the random intercept binary model the weights have a modest effect on sampling variance, which is even reduced for the estimator of the intercept when the design is informative at both levels; moreover, scaling seems to be important in reducing the variance in small samples.

Table 3: Simulation standard deviations of the unscaled weighted point estimators of the intercept and of the second level standard deviation (on 1000 replicates) and corresponding bootstrap estimates (on 150 replicates, with 200 bootstrap samples each) for the design informative at both levels

Sampling design Inform. both levels	*Type of boot.*	β_σ			ω_σ		
		Simul. s.d.	*Boot. estim.*	*Relative error*	*Simul. s.d.*	*Boot. estim.*	*Relative error*
Fixed size $n_j = 38$	FULL	0.193	0.179	-7.3%	0.137	0.116	-15.3%
Fixed size $n_j = 38$	PSU	0.193	0.167	-13.5%	0.137	0.105	-23.4%
Prop. size $n_j = 0.4N_j$	FULL	0.190	0.189	-0.5%	0.142	0.141	-0.7%
Prop. size $n_j = 0.4N_j$	PSU	0.190	0.177	-6.8%	0.142	0.114	-19.7%
Fixed size $n_j = 9$	FULL	0.258	0.369	43.0%	0.289	0.375	29.8%
Fixed size $n_j = 9$	PSU	0.258	0.251	-2.7%	0.289	0.254	-12.1%
Prop. size $n_j = 0.1N_j$	FULL	0.294	0.449	52.7%	0.317	0.440	38.8%
Prop. size $n_j = 0.1N_j$	PSU	0.294	0.258	-12.2%	0.317	0.256	-19.2%

4.2.5 Bootstrap variance estimation

The bootstrap procedure described in Section 3.4 has been applied to estimate the sampling standard deviations of the weighted point estimators of β_σ and ω_σ. We limited the analysis to unscaled weighted estimators and to designs that are informative at both levels. Two different types of bootstrap were implemented (see Section 3.4): in the first type (FULL) the artificial population is generated at both levels, i.e. both the elementary units and the clusters are re-sampled; in the second type (PSU) the steps concerning the elementary units are omitted, i.e. only the clusters are re-sampled. Since the sampling fraction of the clusters is quite low (35/300), we expect that FULL and PSU types of bootstrap will produce similar results. Each simulation comprises 150 replications. For every replication the values of the response variable are generated through the two-stage scheme described in Section 4.1 and 200 bootstrap samples are selected. Table 3 reports, for each parameter, the Monte Carlo standard error of the sampling distribution of the unscaled weighted estimator on 1000 replications of the complex design (see Tables 1 and 2), the corresponding average bootstrap estimate and the relative bias.

Due to the extremely long computational time, we limited our experiment to 150 replications of bootstrap procedures based on 200 bootstrap samples. These numbers are clearly not enough to draw firm conclusions, though some hints about the behavior of the bootstrap estimators can still be derived.

As was to be expected, the PSU-type bootstrap estimate is always smaller than the FULL-type, though the differences are sometimes not so relevant. For both types the performance is slightly better for the estimation of the sampling standard deviation of the estimator of β_σ, rather than of ω_σ.

The sample cluster sizes seem to have little effect on the PSU-type bootstrap estimator, which produced a similar underestimation of the true variability both with small cluster sizes ($n_j = 9$ and $n_j = 0.1N_j$) and with large cluster sizes ($n_j = 38$ and $n_j = 0.4N_j$). On the other hand, the FULL-type bootstrap estimator seems to be strongly influenced by the sample cluster sizes, showing an excellent performance with large cluster sizes and a serious positive bias with small cluster sizes. On these grounds it might be advisable to use PSU-type bootstrap when the sample cluster sizes are very small.

5 Conclusions

The wide use of multilevel ordinal and binary models in many fields of application has motivated our study on the effects of complex sampling designs on the fitting of such models. In the paper we demonstrated, by means of simulations, the bias induced by a two-stage complex sampling design on the fitting of a simple multilevel binary model when the clusters and/or the subjects are selected with probabilities that depend on the response variable. The simulation study also showed that in such situations the bias can be reduced in an effective manner by the probability-weighted estimation procedure we developed in the paper, which is easily implemented in the SAS environment. In particular, the scaled version of the weighted estimator allows to obtain, for both fixed and random parameters, a low bias with a modest increase in the sampling variance. Even if weighting is superfluous, the loss of efficiency due to the inclusion of scaled weights seems to be very low.

A drawback of probability-weighted estimation is the need for special procedures to estimate the variability of the estimators. The standard sandwich variance estimator is not immediately implementable in the SAS environment, so we adopted a bootstrap technique, which is conceptually simple and easy to program, though it requires some computational effort. Our very limited simulation study suggests that its performance is good only for large sample cluster sizes; however much more simulations would be needed to fully understand the behavior of the bootstrap estimator.

References

[1] Gibbons RD, Hedeker D (1997) Random effects probit and logistic regression models for three-level data. *Biometrics*, **53**, 1527-1537.

[2] Hedeker D, Gibbons RD (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933-944.

[3] Pfeffermann D (1993) The role of sampling weights when modeling survey data. *International Statistical Review*, **61**(2), 317-337

[4] Pfeffermann D, Skinner CJ, Holmes DJ, Goldstein H, Rasbash J (1998) Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, **161**(1), 23-40.

[5] Hedeker D, Gibbons RD (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933-944.

[6] Särndal C, Swensson B, Wretman J (1992) *Model assisted survey sampling.* New York: Springer-Verlag.

[7] SAS Institute (1999) *SAS/STAT User's Guide Version 8.* Cary: SAS Institute Inc.

[8] Skinner CJ (1989) Domain means, regression and multivariate analysis. In *Analysis of complex surveys* (eds. CJ Skinner, D Holt and TMF Smith), 59-87. Chichester: Wiley.