Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries

Danny Pfeffermann, Bénédicte Terryn and Fernando A.S. Moura

Abstract

This paper considers situations where the target response value is either zero or an observation from a continuous distribution. A typical example analyzed in the paper is the assessment of literacy proficiency with the possible outcome being either zero, indicating illiteracy, or a positive score measuring the level of literacy. Our interest is in how to obtain valid estimates of the average response, or the proportion of positive responses in small areas, for which only small samples or no samples are available. As in other small area estimation problems, the small sample sizes in at least some of the sampled areas and/or the existence of nonsampled areas requires the use of model based methods. Available methods, however, are not suitable for this kind of data because of the mixed distribution of the responses, having a large peak at zero, juxtaposed to a continuous distribution for the rest of the responses. We develop, therefore, a suitable two-part random effects model and show how to fit the model and assess its goodness of fit, and how to compute the small area estimators of interest and measure their precision. The proposed method is illustrated using simulated data and data obtained from a literacy survey conducted in Cambodia.

Key Words: Credibility intervals; Generalized linear mixed model; Goodness of fit; Linear mixed model; MCMC; Prediction bias; Prediction MSE.

1. Introduction

In this paper we consider situations where the target response value is either zero or an observation from a continuous distribution. A typical example analyzed in the paper is the assessment of literacy proficiency based on a written test with the possible outcome being either zero, indicating illiteracy, or a positive score in a given range measuring the level of literacy. Another example is the consumption of illicit drugs (or certain food items), where a zero value indicates "no consumption", whereas a positive outcome measures the amount consumed. Our interest lies in how to obtain valid estimates of the average response (average literacy level in our example), or the proportion of positive responses (proportion of literate people), in small areas for which only small samples or no samples are available. As in other small area estimation problems, the small sample sizes within the sampled areas and the existence of nonsampled areas requires the use of model based methods.

We propose the use of a two-part random effects model and show how to fit the model and assess its goodness of fit, and how to obtain the small area estimates of interest and measure their precision. The first part of the model specifies the probability of a zero score. The second part specifies the distribution of the positive scores. Although the model is not new and is used in other applications, (see, e.g., Olsen and Schafer 2001 and the discussion and references in that paper), to the best of our knowledge this kind of mixed distribution has not been considered before in the small area estimation literature. Notice that the zero scores in our application are "structural" (true) zeroes. There exists a related body of literature that handles excess of zeros in count data, which may arise from a combination of overdispersion or true zero inflation. Zero inflated data are data that have a larger proportion of zeros than expected from pure count (Poisson) data. See, e.g., Barry and Welsh (2002).

The first part of our model is the logistic function, used to model the probability of a positive score. The second part is a linear model with normal error terms fitted to the non-zero responses. Both models include individual and area level covariates, as well as area random effects that account for variations not explained by the covariates. The model allows for correlations between the corresponding random effects of the two parts and is fitted by application of Markov Chain Monte Carlo (MCMC) simulations.

The two-part model is fitted to data collected as part of the national literacy household survey carried out in Cambodia in 1999, known as the ‘Assessment of the Functional Literacy Levels of the Adult Population’. Figure 1 displays the histogram of the literacy scores observed for this survey. In this application we produce small area estimates for districts of residence and nested villages, requiring the use of a two-part three-level random effects model. We assess the goodness of fit of the model by use of simple descriptive statistics and by simulating data from the model. The use of simulations enables also to compare the results of fitting the ‘full’ two-part model with results obtained by fitting the two parts of the model separately.

1. Danny Pfeffermann, Hebrew University of Jerusalem, Israël, and Southampton Statistical Sciences Research Institute, UK. E-mail: msdanny@huji.ac.il; Bénédicte Terryn, Department for International Development (DFID), UK; Fernando A.S. Moura, Federal University of Rio de Janeiro, Brazil.
without accounting for the correlations between the random effects in the two parts. Another comparison of interest is to results obtained when ignoring the special nature of the data and fitting the linear part to all the responses, ignoring the existence of many zero scores.

In order to facilitate the presentation and discussion in the rest of the paper, we consider literacy scores measured for individuals residing in villages nested in districts, but as noted above, the model considered in this paper can be used for many other important applications.

2. Model and small area predictors

2.1 The two-part model

Let $y$ define the response (literacy test score in our application) and $R$ the covariates and random effects. Then,

$$E(y| R = r) = E(y | R = r, y = 0) \Pr(y = 0 | R = r)$$

$$+ E(y | R = r, y > 0) \Pr(y > 0 | R = r)$$

$$= E(y | R = r, y > 0) \Pr(y > 0 | R = r). \quad (1)$$

For the small area estimation problem considered in this paper we apply a nested three-level model with districts of residence defining the first level, villages defining the second level and individuals defining the third level. For individual $k$ residing in village $j$ of district $i$, with covariates and random effects $R_{ijk} = r$, we have therefore the relationship,

$$E(y_{ijk} | R_{ijk} = r) =$$

$$E(y_{ijk} | R_{ijk} = r, y_{ijk} > 0) \Pr(y_{ijk} > 0 | R_{ijk} = r). \quad (2)$$

In what follows we model the two parts in the right hand side of (2). For individuals with positive responses we assume the ‘linear mixed model’,

$$y_{ijk} = x'_{ijk} \beta + u_i + v_{ij} + e_{ijk};$$

$$u_i \sim N(0, \sigma_u^2); v_{ij} \sim N(0, \sigma_v^2); e_{ijk} \sim N(0, \sigma_e^2). \quad (3)$$

where $x_{ijk}$ represents individual and area level values of covariates, $u_i$ is a random district effect and $v_{ij}$ is a nested random village effect. The random effects $u_i$ and $v_{ij}$, and the residual terms $e_{ijk}$ are assumed to be mutually independent between and within the districts and villages. They account for the variation of the individual scores not explained by the covariates, and define the correlations between the scores of individuals residing in the same village and the correlations between the scores of individuals residing in the same district but in different villages.

$$\text{Corr}(y_{ijk}, y_{ij'}k') =$$

$$\begin{cases}
\frac{(\sigma_u^2 + \sigma_v^2)}{\sigma_v^2} & \text{if } i = i', j = j', k \neq k' \\
\frac{\sigma_u^2}{\sigma_v^2} & \text{if } i = i', j \neq j' \\
0 & \text{if } i \neq i'.
\end{cases} \quad (4)$$

For the probabilities of positive responses (the second part of Equation (2)), we assume the ‘generalized linear mixed model’,

$$p_{ijk} = \Pr(y_{ijk} > 0 | x^*_{ijk}, u^*_i, v^*_{ij}) = \frac{\exp(x^*_{ijk} \gamma + u^*_i + v^*_{ij})}{1 + \exp(x^*_{ijk} \gamma + u^*_i + v^*_{ij})}$$

$$u^*_i \sim N(0, \sigma_u^2); v^*_i \sim N(0, \sigma_v^2). \quad (5)$$

implying that $\logit(p_{ijk}) = \log(p_{ijk} / (1 - p_{ijk})) = x^*_{ijk} \gamma + u^*_i + v^*_{ij}$. Here again, $u^*_i$ and $v^*_i$ represent independent random district and village effects not accounted for by the covariates $x^*_{ijk}$. Notice that the covariates $x^*_{ijk}$ in Equation (3) and the covariates $x^*_{ijk}$ in Equation (5) may differ, see the empirical study in Section 4.

Remark 1. One could argue that the mixed linear model (3) with the added normality assumptions implies a corresponding probit model for the probabilities $p_{ijk}$. This, however, is not true since the model (3) is only assumed for the positive scores. It follows that the parameters of the two models can be assumed to be distinct in the sense of Rubin (1976).

We allow for nonzero correlations between the district random effects in the two parts, and similarly for the village random effects. This is a reasonable assumption since it can be expected that for given values of the covariates, an individual residing in an area characterized by high literacy scores will have a higher probability of a positive score than an individual residing in an area with low scores. The magnitude of these correlations and the importance of

Figure 1 Histogram of literacy scores in the national literacy survey in Cambodia, 1999

Statistics Canada, Catalogue No. 12-001-X
accounting for them when fitting the model depends on the prediction power of the covariates available for the two parts of the model, or alternatively, on the variances of the random effects, (the higher the prediction power of the covariates, the lower are the variances). The correlations are modelled by assuming,

\[ u_i^* | u_i \sim N(K_i u_i, \sigma_{u_i}^2); v_{ij}^* | v_{ij} \sim N(K_i v_{ij}, \sigma_{v_{ij}}^2). \]

(6)

Figure 2 provides supporting evidence for this proposition using the sample data from the center of Cambodia that is used for the empirical study in Section 4. (The empirical correlations between the variables measured on the two axes are 0.25 for villages and 0.38 for districts.)

2.2 Parameters of interest and predictors

For village \((i, j)\) of size \(N_{ij}\), the small area parameters of interest are the true mean of the literacy scores, \(\bar{y}_{ij} = \sum_{k=1}^{S_j} y_{ijk} / N_j\), and the proportion of positive scores, \(P_{ij} = \sum_{k=1}^{S_j} I(y_{ijk} > 0) / N_j\), where \(I(y_{ijk} > 0) = 1\) if \(y_{ijk} > 0\) and is 0 otherwise. Notice that the means are computed over all the individuals, including individuals with zero scores.

Under the model (2), the mean is predicted as,

\[ \hat{y}_{ij} = \frac{\sum_{k=0}^{S_j} y_{ijk} + \sum_{k \in S_j} \hat{y}_{ijk}}{N_j}, \]

(7)

where \(S_j\) defines the sample from village \((i, j)\). By (3) and (5), the missing scores can be predicted under the frequentist approach as,

\[ \hat{y}_{ijk} = \frac{\exp(x_{ijk}^* \hat{\beta} + \hat{u}_i + \hat{v}_j)}{1 + \exp(x_{ijk}^* \hat{\beta} + \hat{u}_i + \hat{v}_j)} \times [x_{ijk}^* \hat{\beta} + \hat{u}_i + \hat{v}_j], \]

(8)

where \(\hat{\beta}, \hat{\gamma}, \hat{u}_i, \hat{v}_j, \hat{u}_i^*, \hat{v}_j^*\) define appropriate sample estimates, see next section. One could add an estimate \(\hat{\epsilon}_{ijk}\) to the estimated mean, \((x_{ijk}^* \hat{\beta} + \hat{u}_i + \hat{v}_j)\), obtained either by drawing from the \(N(0, \sigma^2_{\epsilon})\) distribution, or by selecting at random an estimated residual, \(\hat{\epsilon}_{ijk} = (y_{ij,k} - x_{ijk}^* \hat{\beta} - \hat{u}_i - \hat{v}_j)\) from the estimated residuals computed for the sampled individuals. Adding estimates \(\hat{\epsilon}_{ijk}\) to the estimated mean values reflects more closely the variability of the positive responses. Under the Bayesian approach, the missing scores are predicted by drawing at random from their predictive distribution, see next section.

By (5), the proportion \(P_{ij}\) is predicted under the frequentist approach as,

\[ \hat{P}_{ij} = \frac{1}{N_j} \left[ \sum_{k \in S_j} I(y_{ijk} > 0) + \sum_{k \in S_j} \hat{y}_{ijk} \right], \]

(9)

where

\[ \hat{y}_{ijk} = \frac{\exp(x_{ijk}^* \hat{\beta} + \hat{u}_i + \hat{v}_j)}{1 + \exp(x_{ijk}^* \hat{\beta} + \hat{u}_i + \hat{v}_j)}. \]

A Bayesian solution consists of predicting the indicators \(I(y_{ijk} > 0)\) by drawing at random from their predictive distribution.

The district means and proportions are predicted analogously, which is the same as computing the weighted average of the corresponding village predictors, with the weights defined by the relative village sizes.

Remark 2. The computation of the predictor defined by (7) and (8) requires knowledge of the covariates \(x_{ijk}, x_{ijk}^*\) for every unit in the population. Similarly, the computation of the predictor in (9) requires knowledge of the covariates \(x_{ijk}^*\) for every unit in the population. This is generally true for all generalized linear mixed models. Information on the auxiliary covariate variables is often obtained from censuses or other administrative records. In the absence of such information, the missing covariates can be imputed by drawing at random from their estimated parametric distribution or empirical distribution.

Figure 2 Proportion of positive scores by average of positive scores for districts and villages in center of Cambodia. National literacy survey, 1999
3. **Inference**

The use of the small area predictors defined by (7)-(9) requires estimating the fixed parameters (hyperparameters) \((\beta, \sigma_u^2, \sigma_v^2, \sigma_{\epsilon}^2)\) of the linear part (Equation 3), the fixed parameters \((\gamma, K_u, K_v, \sigma_{\epsilon_{u|\beta}}^2, \sigma_{\epsilon_{v|\gamma}}^2)\) of the logistic part (Equations 5, 6), and predicting the random effects \(\{u_{ij}, v_{ij}; u'_{ij}, v'_{ij}\}\). Methods for estimating fixed and random effects when fitting linear mixed models, or generalized linear mixed models alone, have been developed over the last two decades under both the frequentist and the Bayesian paradigms. The use of these methods permits also the computation of estimators of the mean square error (MSE) or the Bayes risk of the small area predictors that account for hyper parameter estimation to correct order. See the book by Rao (2003) and the more recent article by Jiang and Lahiri (2005) for thorough reviews and discussions. However, the two-part model defined by (2)-(6) has not been considered in the small area literature, and in what follows we consider a few possibilities of fitting this model.

### 3.1 Full likelihood based inference

Define, \(I_{ijk} = 1(0)\) if \(Y_{ijk} > 0(= 0)\) and denote, \(r_{ijk} = (x_{ijk}, u_{ijk}, v_{ijk}), r'_{ijk} = (x_{ijk}, u'_{ijk}, v'_{ijk})\). For given vectors \(r_{ijk}, r'_{ijk}\), the likelihood for the two-part model takes the form,

\[
L = \prod_{i, j, k, s} (p_{ijk})^{s_k}[f(y_{ijk} | r_{ijk}, Y_{ijk} > 0)]^{i_k}(1 - p_{ijk})^{(1-l_k)},
\]

where \(s = \cup S_{ij}\) denotes the sample from all the villages, \(p_{ijk}\) is defined by (5) and \(f(y_{ijk} | r_{ijk}, Y_{ijk} > 0)\) is the normal density with mean \((x_{ijk} \hat{\beta} + u_{ijk} + v_{ijk})\) and variance \(\sigma_{\epsilon}^2\) (Equation 3). The use of this likelihood for inference is, however, problematic because the random effects \(\{(u_{ij}, v_{ij}; u'_{ij}, v'_{ij})\}\) are in fact unobservable. One possibility, therefore, is to integrate the likelihood over the joint (normal) distribution of the random effects as defined by (3) (5) and (6), and maximize the integrated likelihood with respect to the fixed (hyper) parameters \((\beta, \sigma_u^2, \sigma_v^2, \sigma_{\epsilon}^2)\) and \((\gamma, K_u, K_v, \sigma_{\epsilon_{u|\beta}}^2, \sigma_{\epsilon_{v|\gamma}}^2)\). Having estimated the fixed parameters, the random effects can be predicted by their expected values given the data (with the maximum likelihood estimates held fixed), which requires another set of integrations. Olsen and Schafer (2001) consider a two-part model for fitting longitudinal data and approximate the integrated likelihood by a high order multivariate Laplace approximation (Raudenbush, Yang and Yosef 2000). The authors calculate empirical Bayes predictors of the random effects by use of importance sampling (Tanner 1996), setting the fixed parameters at their maximum likelihood estimates. The application of this procedure, however, is very complicated computationally, and the mean square estimators of the errors (MSE) of the small area predictors obtained this way fail to account for the variation induced by estimating the fixed parameters. The contribution to the total MSE from estimating the fixed parameters can not be ignored in general, unless the numbers of sampled districts and villages are very large.

### 3.2 Separate model fitting

The idea here is to fit the two parts of the model separately, and then combine the estimates for computing the small area predictor defined by (7) and (8). The predictor in (9) is obtained directly from fitting the second part only. As mentioned earlier, the fitting of the separate parts has been studied extensively in the literature and computer softwares are readily available, particularly for linear mixed models. It is important to note in this regard that under the present two-part model, the predictors (7)-(9) are nonlinear functions of the data and even when the hyper parameters are known, no explicit formulae are available for the prediction MSEs. Estimating the MSE under the frequentist approach with bias of small order requires therefore developing new appropriate approximations or resampling procedures, which in the case of the predictor \(\hat{y}_{ij}\) defined by (7) and (8), account for the correlations between the data in the two parts. This is further complicated by the fact that by fitting the two parts separately, it is not clear how to estimate the coefficients \((K_u, K_v)\) defining the correlations between the random effects in the two parts (Equation 6). A Jackknife procedure for estimating the prediction MSE of the predictor in (9) under separate model fitting has been developed by Jiang, Lahiri and Wan (2002). Bootstrap estimators applicable to this predictor, again under separate model fitting, are studied in Hall and Maiti (2006).

### 3.3 Bayesian inference under the two-part model

The use of Bayesian methods requires specification of prior distributions for the fixed parameters underlying the two-part model (the coefficients \(\beta, \gamma, K_u, K_v\) and the variances \(\sigma_u^2, \sigma_v^2, \sigma_{\epsilon}^2, \sigma_{\epsilon_{u|\beta}}^2, \sigma_{\epsilon_{v|\gamma}}^2\), but with the aid of Markov Chain Monte Carlo (MCMC) simulations, the application of this approach permits sampling from the posterior distribution of the fixed parameters and the random effects, and hence sampling from the predictive distribution of the unobserved responses. Thus, the use of this approach yields the whole posterior distribution of the small area parameters of interest, allowing thereby the computation of correct MSE (posterior variance) measures or confidence (credibility) intervals that account for all the sources of variation. As discussed above, estimation of the prediction MSE under the previous approaches is problematic, particularly with regard to the predictor \(\hat{y}_{ij}\) defined by (7) and (8). Computer software is available to
perform all the necessary computations but it should be noted that with complex models, the computations can be intensive and time consuming.

In the empirical study of this article we followed the Bayesian approach using the WinBUGS software (Spiegelhalter, Thomas and Best 2003). This software is known to be “user friendly”, and based on our past experience it operates very well. Clearly, there are many other software available for MCMC simulations, such as MLwiN (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse, Draper, Langford and Lewis 2002) or R (Development Core Team 2008). WinBUGS implements the MCMC algorithm with the Gibbs sampler (Gelfand and Smith 1990). The Gibbs sampler samples alternately from the conditional distribution of each of the fixed and random parameters (random effects), given the data and the remaining parameters. It defines a Markov chain, which under some regularity conditions converges to a realization from the joint posterior distribution of all the model parameters. Thus, at the end of the sampling process (upon convergence), the algorithm produces a (single) realization of each of the fixed and random parameters from their joint posterior distribution given the data. The realizations are denoted below by a tilde above the symbols. Realizations \( \tilde{y}_{jk} \) from the posterior distribution of \( y_{jk} \) are obtained by randomly drawing \( \tilde{I}_{jk} = 1 \) (or 0) with probabilities \( \tilde{p}_{jk} = \exp(x_{jk}^T \tilde{\gamma} + \tilde{u}_j + \tilde{v}_y) \times [1 + \exp(x_{jk}^T \tilde{\gamma} + \tilde{u}_j + \tilde{v}_y)]^{-1} \), and defining,

\[
\tilde{y}_{jk} = (x_{jk}^T \tilde{\beta} + \tilde{u}_j + \tilde{v}_y + \tilde{v}_{ijk}) \times \tilde{I}_{jk}.
\] (11)

Substituting \( \tilde{y}_{jk} \) for \( y_{jk} \) in (7) and \( \tilde{I}_{jk} \) for \( I_{jk} \) in (9) yields a single sampled value of the mean \( \tilde{Y}_j \) and the proportion \( \tilde{P}_j \) from their respective posterior distributions, for every village \( (i, j) \). Repeating the same process independently a large number of times (using parallel chains, see below) yields an empirical approximation to the posterior distribution of the mean and the proportion. The true village means are then predicted by averaging the corresponding sampled values in all the chains and similarly for the village proportions. The MSE (Bayes risk) is estimated by computing the empirical variance of the sampled values. Credibility (confidence) intervals with coverage rates of \( (1 - \alpha) \) are defined by the \( \alpha/2 \) and \( (1 - \alpha/2) \) level quantiles of the empirical posterior distribution. The same procedure is applied for predicting the district means and proportions, and the corresponding for computing prediction variance and credibility intervals.

In practice, the use of parallel chains for producing independent realizations from the posterior distributions is often too time consuming, in which case the samples can be generated from a single long chain or a few chains, but selecting only every \( r \)th sampled value (after convergence), thus reducing as much as possible the correlations between adjacent sampled values.

4. Empirical results

4.1 Data and model

We use data from the 1999 survey, ‘Assessment of the Functional Literacy Levels of the Adult Population’ in Cambodia for the empirical illustrations. This is a household survey, interviewing 6,548 adults and administering a literacy test consisting of 20 tasks in the Khmer language, with scores ranging from 0 to 100 (see Figure 1 in the introduction). The survey used a stratified multi-stage sampling design with the strata defined by the 24 provinces that comprise the country. Each of the provinces is divided into districts, and about half of them were selected to the sample (a total of 96 districts out of the 184 districts in the country). Two communes were sampled from each of the selected districts and other than in a few cases, three villages were selected from each of the sampled communes. Finally, households were sampled in each village and one adult selected from each household, alternating according to age and sex. The sampling design at each stage was systematic sampling. The number of households selected in each village was the same for all the villages belonging to the same province. The total province sample sizes were allocated proportionally to the province population sizes.

The small areas of interest are the districts and villages. In the present study we restrict to the 50 rural districts sampled in provinces located in the center of the country, for which the same model is expected to hold. In these 50 districts 5 districts had samples of 20 adults or less, and the remaining 45 districts had samples of 41 to 120 adults. The number of villages in the reduced data set is 286, with 47 villages having samples of 9 or less adults and 193 villages having samples of 10 to 19 adults. The total number of adults in the sample is \( n = 4,028 \).

Table 1 shows the results obtained when fitting the full two part model to the sample data, using the Bayesian methodology and software described in Section 3.3. The covariate (regressor) variables in the two models have been selected by application of some standard model selection procedures. All the covariates except for age, education and household size are dummy variables, taking the value 1 when the variable definition is satisfied. We used normal prior distributions with large variances for the elements of the vector coefficients \( \beta, \gamma \), and uniform priors with large (but finite) range for the standard deviations underlying the two parts of the model and the coefficients \( K_u \) and \( K_v \) in Equation 6. By default, WinBUGS automatically selects the method of sampling from the conditional distribution of
each of the fixed and random parameters when applying the Gibbs sampler. Notice that the conditional distributions don’t have a closed-form under the present full model. The software selects an acceptance/rejection method for the logistic part, and slice sampling (Neal 2000) for most of the other parameters and random effects.

For the MCMC simulations we generated a chain of length 50,000, discarded the first 5,000 sampled values as “burn in”, and then thinned the chain by taking every 150th sample value. Discarding the first 5,000 sampled values was found sufficient to guarantee the convergence of the chain, using some informal commonly used graphical techniques. These include comparing the histograms of the posterior distributions of the various parameters based on different sub-sequences of the chain, inspecting the traces of several chains simulated in parallel, each with different starting values to check for stabilization of the chain, and plotting the autocorrelations of the sampled values to verify independence after appropriate thinning. See Gamerman and Lopes (2006) for further discussion and illustrations, including more formal tests of convergence. Note also that the simulation results in Section 4, using the model fitted to the real data and generating a separate chain of length 50,000 for each simulation and discarding the first 5,000 values as “burn in” yield very satisfactory results, thus providing another indication for the convergence of the chain after the first 5,000 values.

The estimated K-coefficients and variances of the random effects imply, Corr \((\hat{u}_i, \hat{u}_j) = 0.45; \text{Corr} (\hat{v}_i, \hat{v}_j) = 0.21. Interestingly, the correlations are close to the empirical correlations reported at the end of Section 2.1, using the raw means.

The main results emerging from Table 1 can be summarized as follows. All the regressor coefficients are highly significant (based on standard t-tests) and generally have anticipated signs. Other variables considered for inclusion in the two models were found to be nonsignificant. The variances of the random effects are highly significant in both models, indicating their contribution in explaining the variation of the scores, or the probabilities of positive scores, not explained by the covariates included in the two models.

As a further diagnostic for the logistic mixed model we show in Figure 3 a scatter plot of the observed proportions of positive scores \((I_{ijk} = 1)\) against the average of the predicted probabilities of positive scores under the model, in groups of 50 individuals defined by the ordered values of the predicted probabilities. The plotted values are almost on a straight line, showing a good fit. Figure 4 shows a histogram of the estimated standardized residuals of the mixed linear part, \(\hat{z}_{ijk} = \hat{e}_{ijk}/SD(\hat{e}_{ijk}) = (y_{ijk} - x'_{ijk} \hat{\beta} - \hat{u}_i - \hat{v}_j)/SD(\hat{e}_{ijk})\), where SD(\(\hat{e}_{ijk}\)) is the empirical standard deviation of the estimated residuals. Although not a ‘perfect’ bell shape, the histogram does not indicate severe divergence from a normal distribution.

Table 1
Estimated parameters and standard errors (Std Err.) when fitting the two-part model

<table>
<thead>
<tr>
<th>Regressors</th>
<th>Linear part</th>
<th></th>
<th>Logistic part</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std Err.</td>
<td>Estimate</td>
<td>Std Err.</td>
</tr>
<tr>
<td>Constant</td>
<td>(\hat{\beta}_0 = 6.90)</td>
<td>4.00</td>
<td>(\hat{\gamma}_0 = -6.48)</td>
<td>0.58</td>
</tr>
<tr>
<td>Years at school</td>
<td>(\hat{\beta}_1 = 7.28)</td>
<td>0.53</td>
<td>(\hat{\gamma}_1 = 2.16)</td>
<td>0.12</td>
</tr>
<tr>
<td>Years at school(^2)</td>
<td>(\hat{\beta}_2 = -0.24)</td>
<td>0.05</td>
<td>(\hat{\gamma}_2 = -0.13)</td>
<td>0.01</td>
</tr>
<tr>
<td>Attended literacy program</td>
<td>-</td>
<td>-</td>
<td>(\hat{\gamma}_3 = 2.44)</td>
<td>0.27</td>
</tr>
<tr>
<td>Helped by interviewer</td>
<td>-</td>
<td>-</td>
<td>(\hat{\gamma}_4 = 2.00)</td>
<td>0.17</td>
</tr>
<tr>
<td>Low income</td>
<td>(\hat{\beta}_5 = -2.61)</td>
<td>0.88</td>
<td>(\hat{\gamma}_5 = -0.35)</td>
<td>0.14</td>
</tr>
<tr>
<td>Civil servant/professional</td>
<td>(\hat{\beta}_6 = 13.91)</td>
<td>1.89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gender (1 for female)</td>
<td>(\hat{\beta}_7 = -1.60)</td>
<td>0.81</td>
<td>(\hat{\gamma}_7 = -0.59)</td>
<td>0.14</td>
</tr>
<tr>
<td>Household size (adults)</td>
<td>(\hat{\beta}_8 = 0.94)</td>
<td>0.29</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age</td>
<td>(\hat{\beta}_9 = 0.84)</td>
<td>0.16</td>
<td>(\hat{\gamma}_9 = 0.14)</td>
<td>0.02</td>
</tr>
<tr>
<td>Age(^2)</td>
<td>(\hat{\beta}_{10} = -0.01)</td>
<td>0.002</td>
<td>(\hat{\gamma}_{10} = -0.002)</td>
<td>0.00</td>
</tr>
<tr>
<td>Variances</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Districts</td>
<td>(\hat{\sigma}_u^2 = 66.31)</td>
<td>16.72</td>
<td>(\hat{\sigma}_{u'}^2 = 1.28)</td>
<td>0.34</td>
</tr>
<tr>
<td>Between Villages</td>
<td>(\hat{\sigma}_v^2 = 66.58)</td>
<td>10.45</td>
<td>(\hat{\sigma}_{v'}^2 = 0.86)</td>
<td>0.19</td>
</tr>
<tr>
<td>Residual</td>
<td>(\hat{\sigma}_x^2 = 322.0)</td>
<td>10.12</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\(K\)-Coefficients (Equation 6)*

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>District random effects</td>
<td>(\hat{K}_u = 0.06)</td>
<td>0.02</td>
</tr>
<tr>
<td>Village random effects</td>
<td>(\hat{K}_v = 0.02)</td>
<td>0.01</td>
</tr>
</tbody>
</table>
As a final assessment of the goodness of fit of the two part model, we generated 200 new data sets of size \( n = 4,028 \) from the estimated two-part model of Table 1, using the same covariates as for the original sample. The test scores where generated by generating random effects and residuals \((u_i, v_{ij}, \varepsilon_{ijk})\) with estimated variances \((\hat{\sigma}^2_u, \hat{\sigma}^2_v, \hat{\sigma}^2_{\varepsilon})\) (Equation 3), generating random effects \((u_i^*, v_{ij}^*)\) using Equation (6) with estimated coefficients and variances \((\hat{K}_u, \hat{\sigma}^2_{u|u^*}, \hat{K}_v, \hat{\sigma}^2_{v|v^*})\), drawing at random 1 or 0 with probabilities \(\Pr(I_{ijk} = 1) = \exp(x_{ijk}^* \gamma + u_i^* + v_{ij}^*)/[1 + \exp(x_{ijk}^* \gamma + u_i^* + v_{ij}^*)]\) (Equation 5), and in the case of 1, generating the nonzero scores \(y_{ijk} = x_{ijk}^* \beta + u_i + v_{ij} + \varepsilon_{ijk}\) (Equation 3). The variance \(\hat{\sigma}^2_{u|u^*}\) was computed as \(\hat{\sigma}^2_{u|u^*} = (\hat{\sigma}^2_u - \hat{K}_u \hat{\sigma}^2_u)\), and similarly for \(\hat{\sigma}^2_{v|v^*}\) (Equation 6). Next we calculated for each data set the score means and proportions for each village and district and used them to compute empirical confidence intervals based on the 200 means and proportions. Table 2 shows the proportions of times that the empirical confidence intervals (C.I.) contain the corresponding actual sample values in the Cambodia survey.

The results in Table 2 show very close coverage rates to the nominal values for the villages, but under-coverage of up to 10% for the districts, which is probably explained by the fact that the latter rates are based on only 50 districts.

### 4.2 Simulation study

The purpose of the simulation experiment is to study the effectiveness of the two-part model for producing small area predictors and associated measures of prediction errors. The simulation experiment enables also to compare the results obtained under this model with results obtained when fitting the two parts of the model separately, ignoring the correlations between the corresponding random effects in the two parts, and with the results obtained when fitting a linear mixed model to all the responses, ignoring the accumulation of zero scores. To this end, we generated 300 new populations of \(N = 4,028\) scores and 300 new samples of size \(n = 1,026\), similar to the generation of the data sets used for the computation of the confidence intervals in Table 2, but from a model with fewer regressors than in the model shown in Table 1. In the logistic part we included 4 regressors: ‘number of years at school’, ‘attendance of a literacy programme’, ‘helped by the interviewer’ and ‘having low income’. In the linear part we included 5 regressors: ‘number of years at school’, ‘gender’, ‘household size’, ‘age’, and ‘age^2’. In order to set parameter values, we fitted separately the linear part and the logistic part with the fewer regressors to the original sample data. The correlations between the random effects of the logistic and the linear parts were set to 0.5 at both the district and the village level.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Proportions of times that the empirical confidence intervals contain the actual sample means and proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical 90% C.I.</td>
</tr>
<tr>
<td>Small areas</td>
<td>Districts</td>
</tr>
<tr>
<td>Number of areas</td>
<td>50</td>
</tr>
<tr>
<td>% Coverage of proportions</td>
<td>80%</td>
</tr>
<tr>
<td>% Coverage of means</td>
<td>88%</td>
</tr>
</tbody>
</table>
The district and village means (proportions) in the simulated populations were taken as the true district and village means (proportions), thus allowing us to assess the performance of the various predictors. As noted in Remark 2 in Section 2.2, the prediction of small areas means and proportions under the two-part model requires knowledge of the covariates for all the population units. This requirement was satisfied in the simulation study since the simulated populations use the regressors of the original sample of the 4,028 individuals. In order to specify sampled values for the regressor variables, we sampled 1,026 individuals and used the sampled regressors for all the 300 samples. Half of the individuals in half of the 286 villages were included in the sample, except for villages with fewer than 5 adults in the original dataset, where all the individuals were sampled. This minimum size criterion was applied in order to avoid computational problems when running the simulations (see Section 4.3). The sample contained individuals from all the 50 rural districts, with 1 district having a sample of size 4, 4 districts having a sample of size 9, 17 districts having samples of size 15 ≤ nj ≤ 20, and the remaining 28 districts having samples of size 21 ≤ nj ≤ 30. As mentioned above, the sample contained individuals from half of the 286 villages, with 29 villages having samples of size 2 ≤ n_j ≤ 5, 109 villages having samples of size 6 ≤ n_j ≤ 10, and 5 villages having samples of size 11 ≤ n_j ≤ 19.

The results of the simulation study are shown in Tables 3 and 4 and in Figures 5-6. Table 3 shows the mean estimates of the model coefficients and the root mean square errors (RMSE) over the 300 simulations, as obtained when fitting the three models to the sample data; A-the full two-part model that accounts for the correlations between the district and village random effects in the two parts of the model, B-the two part model that ignores the correlations between the district and village random effects in the two parts, that is, when fitting the two parts separately, and C-the linear mixed model defined by (3) but fitted to all the responses, including the zero scores. This model ignores the accumulation of zero scores, but in order to make it more comparable to the two part model, we included in this model all the regressors included in either the logistic or the linear part of the two-part model. The linear mixed model can practically only be used for predicting the district and village means. For comparability reasons we fitted all the three models using the WinBUGS software (thus following the Bayesian paradigm), but it is important to mention that fitting the models B and C using the MLwiN software (Rasbash et al. 2002), which is much faster, yields very similar results.

Table 3 exhibits only minor differences between the mean estimates and RMSEs when fitting the full model or when fitting the two parts separately. For the linear part the mean estimates are very close to the corresponding true coefficients, indicating lack of bias. For the logistic part the mean estimates are again close to the true coefficients although the estimated biases are statistically significant based on the conventional t-statistic. The fact that the RMSEs are similar when fitting the full model and when fitting the two parts separately suggests that under the present simulation set-up, accounting for the correlations between the random effects in the two parts does not improve the estimation of the model regression coefficients. In contrast, the results in Table 4 reveal much smaller biases and RMSEs when estimating the variances of the logistic model by fitting the full model, although the estimation of the “between villages” variance is still highly biased. The estimation of the correlations between the random effects of the two parts is satisfactory. Finally, as indicated by both tables, fitting the mixed linear model, ignoring the accumulation of zeroes generally yields highly biased estimators and consequently large RMSEs, which of course is not surprising.

Figure 5 shows the bias and RMSE when predicting the true district and village means and proportions under the three models. Let \( \hat{U}_d \) represent any of the predictors under the three models (means or proportions) for a given area \( a \) as obtained in simulation \( r \), and denote by \( U_{d,a} \) the corresponding true predicted value. The bias and RMSE were calculated as,

\[
\text{Bias}_a = \frac{\sum_{r=1}^{300} (\hat{U}_d - U_{d,a})}{300} ;
\]

\[
\text{RMSE}_a = \left[ \frac{\sum_{r=1}^{300} (\hat{U}_d - U_{d,a})^2}{300} \right]^{1/2} .
\]  

(12)

The figures pertaining to villages are based on the 273 villages (out of the 286) where sampling took place. (As mentioned before, all the individuals in villages with fewer than 5 adults in the original dataset were included in the sample.)

The clear conclusion from Figures 5a, 5c, 5e and 5g is that the use of the mixed linear model alone for predicting the district and village means yields biased predictors in both sampled and nonsampled areas, and hence large RMSEs. Note, however, that the RMSEs of the predictors produced under the linear model for villages without samples are similar to the RMSEs obtained under the two-part model. This outcome is probably explained by the fact that the mixed linear model is much simpler and depends on fewer parameters than the two part model, resulting in smaller prediction variances in villages with no samples than the prediction variances of the two-part model predictors. Figures (5a)-(5d) show that the predictors produced under the two-part model, whether fitted jointly or separately are basically unbiased, despite the bias in the estimation of some of the parameters of the logistic part noticed in Tables 3 and 4. Figures (5e)-(5h) don’t show any
appreciable difference in the RMSE between the use of the full model or by fitting the two parts separately, which was noted also in Tables 3 and 4.

Figure 6 shows the percentage of times that 95% credibility intervals, produced under the three models, cover the true district or village means and proportions. See Section 3.3 for the construction of credibility interval boundaries when using MCMC simulations. The prominent conclusion emerging from Figure 6 is that ignoring the accumulation of zeroes and fitting the linear mixed model alone yields for most areas coverage rates for the true area means that are very different from the nominal 95% rate, with particularly low rates for villages with samples. The fitting of the full model yields somewhat better coverage rates for the district means than the fitting of the two parts separately, but the coverage rates of the district proportions are similar under the two methods. There seems to be little difference in the credibility intervals for the village means when fitting the full model or the two-parts separately, but it is interesting to note that the use of the full model yields better coverage rates in 77 per cent of the villages, whereas fitting the two parts separately yields better coverage rates in only 15 per cent of the villages. In the remaining villages the use of the two methods yields the same coverage rates. In the case of the village proportions, the two methods yield similar credibility intervals, except in a few cases where the use of the full model is seen to be generally better.

Table 3
Means and RMSE of estimators of model coefficients under the three models

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>True value</th>
<th>Full model</th>
<th>Separate fit</th>
<th>Linear model</th>
<th>Full model</th>
<th>Separate fit</th>
<th>Linear model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Simulation mean</td>
<td>Simulation RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear part</td>
<td></td>
<td>Full model</td>
<td>Separate fit</td>
<td>Linear model</td>
<td>Full model</td>
<td>Separate fit</td>
<td>Linear model</td>
</tr>
<tr>
<td>β₀</td>
<td>9.38</td>
<td>8.83</td>
<td>9.73</td>
<td>1.90</td>
<td>6.95</td>
<td>6.95</td>
<td>9.21</td>
</tr>
<tr>
<td>β₁</td>
<td>4.97</td>
<td>4.97</td>
<td>4.87</td>
<td>12.59</td>
<td>0.32</td>
<td>0.33</td>
<td>7.63</td>
</tr>
<tr>
<td>β₂</td>
<td>-1.65</td>
<td>-1.61</td>
<td>-1.58</td>
<td>-3.24</td>
<td>2.05</td>
<td>2.05</td>
<td>2.27</td>
</tr>
<tr>
<td>β₃</td>
<td>1.02</td>
<td>1.05</td>
<td>1.05</td>
<td>1.75</td>
<td>0.57</td>
<td>0.57</td>
<td>0.86</td>
</tr>
<tr>
<td>β₄</td>
<td>0.94</td>
<td>0.97</td>
<td>0.96</td>
<td>1.51</td>
<td>0.27</td>
<td>0.26</td>
<td>0.60</td>
</tr>
<tr>
<td>β₅</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Logistic part</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₀</td>
<td>-4.09</td>
<td>-4.38</td>
<td>-4.38</td>
<td>-</td>
<td>0.58</td>
<td>0.59</td>
<td>-</td>
</tr>
<tr>
<td>γ₁</td>
<td>1.63</td>
<td>1.73</td>
<td>1.73</td>
<td>-</td>
<td>0.18</td>
<td>0.19</td>
<td>-</td>
</tr>
<tr>
<td>γ₂</td>
<td>1.98</td>
<td>2.13</td>
<td>2.13</td>
<td>2.55*</td>
<td>0.41</td>
<td>0.42</td>
<td>7.33*</td>
</tr>
<tr>
<td>γ₃</td>
<td>2.06</td>
<td>2.41</td>
<td>2.41</td>
<td>2.05*</td>
<td>0.65</td>
<td>0.65</td>
<td>2.64*</td>
</tr>
<tr>
<td>γ₄</td>
<td>-0.35</td>
<td>-0.34</td>
<td>-0.34</td>
<td>0.38*</td>
<td>0.30</td>
<td>0.30</td>
<td>1.37*</td>
</tr>
</tbody>
</table>

*Estimates obtained when including these regressors in the linear model.

Table 4
Means and RMSE of estimators of model variances and correlations under the three models

<table>
<thead>
<tr>
<th></th>
<th>True value</th>
<th>Full model</th>
<th>Separate fit</th>
<th>Linear model</th>
<th>Full model</th>
<th>Separate fit</th>
<th>Linear model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Simulation mean</td>
<td>Simulation RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variances – linear part</td>
<td></td>
<td>Full model</td>
<td>Separate fit</td>
<td>Linear model</td>
<td>Full model</td>
<td>Separate fit</td>
<td>Linear model</td>
</tr>
<tr>
<td>District</td>
<td>60.40</td>
<td>62.23</td>
<td>60.66</td>
<td>103.46</td>
<td>24.52</td>
<td>24.87</td>
<td>46.52</td>
</tr>
<tr>
<td>Village</td>
<td>65.44</td>
<td>70.37</td>
<td>70.36</td>
<td>111.97</td>
<td>24.84</td>
<td>25.75</td>
<td>49.74</td>
</tr>
<tr>
<td>Residual</td>
<td>336.00</td>
<td>338.31</td>
<td>338.61</td>
<td>696.82</td>
<td>23.76</td>
<td>24.04</td>
<td>361.64</td>
</tr>
<tr>
<td>Variances – logistic part</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>District</td>
<td>0.92</td>
<td>1.08</td>
<td>1.50</td>
<td>-</td>
<td>0.61</td>
<td>0.91</td>
<td>-</td>
</tr>
<tr>
<td>Village</td>
<td>0.57</td>
<td>0.91</td>
<td>1.15</td>
<td>-</td>
<td>0.70</td>
<td>0.94</td>
<td>-</td>
</tr>
<tr>
<td>K-factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>District</td>
<td>0.071</td>
<td>0.075</td>
<td>-</td>
<td>-</td>
<td>0.016</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Village</td>
<td>0.054</td>
<td>0.055</td>
<td>-</td>
<td>-</td>
<td>0.012</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Correlations between random effects of the two parts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>District</td>
<td>0.500</td>
<td>0.506</td>
<td>-</td>
<td>-</td>
<td>0.151</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Village</td>
<td>0.500</td>
<td>0.459</td>
<td>-</td>
<td>-</td>
<td>0.148</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
(5a) bias, district means

(5b) bias, district proportions

(5c) bias, village means
Villages ordered by ascending sample size (first half: Villages with no sample)

(5d) bias, village proportions

Districts ordered by ascending sample size

(5e) RMSE, district means

Districts ordered by ascending sample size

(5f) RMSE, district proportions
Figure 5 BIAS and RMSE of predictors of area means and proportions

(5g) RMSE, village means

(5h) RMSE, village proportions

Figure 5 BIAS and RMSE of predictors of area means and proportions

(6a) district means
4.3 Computational issues

As already noted, fitting the full model, accounting for the correlations between the district and village random effects in the two parts is computationally intensive and not always stable. In particular, we encountered severe computation problems when fitting the full model with very small samples from most of the villages. For example, for a sample of 750 individuals from 264 villages, such that almost half of the villages had sample sizes of 1 or 2, the sampled values...
from the posterior distributions generated by the Gibbs sampler were found to be strongly correlated even at very high lags, over 1,000 lags for the village random effects and the correlation between the village random effects in the two parts, and still over 500 lags after tightening the prior distributions, which required extremely long chains to obtain sufficient data for inference. This makes it excessively computer intensive and almost impossible to verify convergence of some of the posterior distributions. For this reason we selected samples of size 1,026 in our simulation study, with at least 2 individuals from every village.

5. Summary

The most important message emerging from this paper is that ignoring the accumulation of zeroes and fitting a linear mixed model to the whole data set can result in highly biased predictors and wrong coverage rates of credibility intervals. Clearly, the magnitude of the bias and the performance of the credibility intervals will depend in this case on the percentage of zero scores. Fitting a two-part model to such data generally yields unbiased predictors and credibility intervals with acceptable coverage rates. Fitting the full two-part model, accounting for the correlations between the random effects of the two parts is the best choice, but it improved the predictions in our simulation study only marginally, despite the use of correlations of 0.5 between the district and village random effects in the two parts.

In this study we used MCMC simulations for fitting the models and computing the small area predictors and their variances. The use of this approach requires specifying prior distributions, which can affect the inference, particularly with a small number of sampled areas even when specifying noninformative priors. See Pfeffermann, Moura and Silva (2006) for recent discussion and illustrations. The other problem with the use of MCMC simulations is that it is very computing intensive. Furthermore, the use of this approach can become unstable if there are only few observations in the sampled areas. An alternative approach is therefore to fit the full two-part model following the frequency approach. Available software include MLwiN (Goldstein 2003) and aML (Lillard and Panis 2003), but the use of these or other softwares requires modifications to the estimation of the prediction variance that account for the errors in the estimation of the fixed model parameters. Resampling methods like the bootstrap or jackknife could be considered for this purpose, but they require new developments appropriate for this model.

Acknowledgements

The work of Bénédicte Terryn was conducted for the most part while at UNESCO Institute for Statistics (UIS), Montreal, Canada. The work of Fernando Moura was funded in part by a research grant from the Brazilian National Council for the Development of Science and Technology. The authors thank an associate editor and two reviewers for their constructive comments and suggestions, and Pedro Silva for helping with the final computations.

References


