Reviewer's report

Title: Fitting multilevel models in complex survey data with design weights: Recommendations

Version: 4 Date: 4 April 2009

Reviewer: Sophia Rabe-Hesketh

Reviewer's report:

Review of revised version of “Multilevel modeling and weighted complex survey data”

The paper has been improved considerably.

Now that the details of the estimation methods have been revealed, I realize that the estimates are sometimes not comparable across software packages, or that the method used within a particular program sometimes is not the one that is recommended in the software manual:

1. It does not make sense to report robust standard errors for Mplus and model-based standard errors for MLwiN and gllamm. They are different concepts, and it does not make sense to present them next to each other in a table as if it made sense to compare them on order to judge the performance of different software packages. It does not matter what the default settings are: Either present robust standard errors for all three packages, or model-based standard errors for all three packages.

2. For MLwiN, the manual does not recommend using MQL. The gold-standard is PQL, or even PQL with bias-correction using parametric bootstrapping. By the way, MQL is not IGLS (p.12, line 10).

3. The gllamm manual clearly states that it is up to the user to determine how many quadrature points are required. The recommendation is to increase the number of quadrature points until the estimates stabilize. The recommendation also is to use adaptive quadrature which is not the default - it requires the adapt option. (Therefore, it’s not clear whether adaptive quadrature was in fact used.) If 16 points were required, report these estimates in the table. (Make sure you use the adapt option - perhaps 8 points will be sufficient if you do,) It is not sufficient to recommend to readers that they should increase the number of quadrature points. A methodological paper must lead by example.

In my last review (point 5 in the author’s response), I pointed out that it is not true that traditional survey approaches “fail to allow analysts to investigate what predicts variation across clusters”. You suggest adding "and examine the amount of between-cluster variance unaccounted for by predictors in the model”. While I agree with the second part of the sentence, the first part of the sentence is still incorrect. It is a common misconception that multilevel models are needed for including cluster-level covariates and cross-level interactions. However, this is
just as easily accomplished in ordinary regression models with standard errors adjusted for clustering. The following statements should therefore be removed from the paper: (1) "However, it fails to allow analysts to investigate what predicts variation across clusters" (p.3, line -6); (2) "analysts need to adopt methods that allow investigations of both individual and cluster level predictors" (P.3, bottom). "In this way, analysts can examine predictors at both the cluster and individual levels" (P.4, line -11).

Specific comments

P.3, line -8: Delete "variable" (the clustered nature is not a variable).

P.4, line 9: Delete "variable" (the design is not a variable).

P.4, line -3: Rewrite "a function of sums across" - the critical point is that the level-1 weights and level-2 weights appear in separate places within that function, so it is not sufficient to know the product of these weights.

P.6, line 2: Delete "Among other issues", or clarify what you mean.

P.7, line -7: "often"? It never will if level 2 is the highest level of the model and if the same scale factor is applied for all units.

P.8, line 10: Rewrite "point estimate variance across clusters". Do you mean the residual between-cluster variance?

P.10, line -9: Should be "adjusted".

P.10, line -7: Rewrite "and do not need nor does NCHS include"

P.12, line 1: I think this is a misunderstanding of Raudenbush & Bryk. I don't think they recommend that level-2 coefficients should always be treated as random (by the way, do you mean level-1 coefficients?).

P.12, line -10: MQL is not IGLS.

P.13, line -7: MLwiN does not use MQL for continuous responses.

P.14, line-5: "empirically describe confidence in the results"? What does that mean?

P.16, line 9-10: What is the "residual standard error variance"? Either "residual variance" (a model parameter), or "standard error" (the imprecision of the estimate of a model parameter).

P.17, line 2: Should be "link function".


P.17, line -7: Replace reference 23 by reference 10.

P.29, SAS code and Stata code: This looks more complicated than necessary
because you are dividing the weights by \((a/b)\), instead of just multiplying them by \(b\) and dividing them by \(a\). For example, it would be much clearer to write:

\[ aw = \text{weight}_i \times \left( \frac{n_j}{\text{sum}w} \right) \]

P.30, the expression for the weights for Method B is incorrect.

Appendix C. Avoid the asterisk \(*\) to denote a product. Use \(\times\) (the "times" symbol). Define the variance components. Perhaps say that in all models, \(\#00\) is the variance of \(u_{0j}\), etc. (or whatever notation you want to use).

Tables: The symbols must match the symbols in the equations in the appendix. (For example, I do not see any \(\#00\), etc., in the equations, but they appear in the tables). For variance components, make your mind up what notation to use \(\#2\) or \(\#\) and define these in the equations.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.