Health and Place 52 (2018) 121–126

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

Does an uneven sample size distribution across settings matter in cross-classified multilevel modeling? Results of a simulation study

Carly E. Milliren[a,b,*], Clare R. Evans[c], Tracy K. Richmond[b,d], Erin C. Dunn[e,f,g]

[a] Center for Applied Pediatric Quality Analytics, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA
[b] Division of Adolescent/Young Adult Medicine, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA
[c] Department of Sociology, University of Oregon, 736 PLC 1291, Eugene, OR 97403, USA
[d] Department of Pediatrics, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA
[e] Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA
[f] Department of Psychiatry, Harvard Medical School, 401 Park Drive, Boston, MA 02215, USA
[g] Stanley Center for Psychiatric Research, The Broad Institute of Harvard and MIT, 75 Ames Street, Cambridge, MA 02142, USA

ARTICLE INFO

Keywords:
Cross-classified multilevel modeling
Contextual effects
Simulation
Schools
Neighborhoods

ABSTRACT

Background: Recent advances in multilevel modeling allow for modeling non-hierarchical levels (e.g., youth in non-nested schools and neighborhoods) using cross-classified multilevel models (CCMM). Current practice is to cluster samples from one context (e.g., schools) and utilize the observations however they are distributed from the second context (e.g., neighborhoods). However, it is unknown whether an uneven distribution of sample size across these contexts leads to incorrect estimates of random effects in CCMMs.

Methods: Using the school and neighborhood data structure in Add Health, we examined the effect of neighborhood sample size imbalance on the estimation of variance parameters in models predicting BMI. We differentially assigned students from a given school to neighborhoods within that school's catchment area using three scenarios of (im)balance. 1000 random datasets were simulated for each of five combinations of school- and neighborhood-level variance and imbalance scenarios, for a total of 15,000 simulated data sets. For each simulation, we calculated 95% CIs for the variance parameters to determine whether the true simulated variance fell within the interval.

Results: Across all simulations, the "true" school and neighborhood variance parameters were estimated 93–96% of the time. Only 5% of models failed to capture neighborhood variance; 6% failed to capture school variance.

Conclusions: These results suggest that there is no systematic bias in the ability of CCMM to capture the true variance parameters regardless of the distribution of students across neighborhoods. Ongoing efforts to use CCMM are warranted and can proceed without concern for the sample imbalance across contexts.

1. Introduction

Multilevel modeling (MLM) has become a staple of social science and public health research, allowing researchers to examine macro-level contextual effects across multiple settings, including students within schools (Munoz and Chang, 2007; Kim and McCarthy, 2006; Sellstrom and Bremberg, 2006), residents within neighborhoods (Tendulkar et al., 2010; Pickett and Pearl, 2001; Leventhal and Brooks-Gunn, 2000), and patients within hospitals (Rice and Alastair 1996). In MLM, both fixed and random effects account for the clustering of individuals within context, while also generating effect estimates for the contexts themselves (Diez-Roux, 2000). For more than two decades, studies using MLM have demonstrated that contexts are important determinants of health and behavior, even after accounting for individual characteristics and composition.

Recently, MLM researchers have begun to recognize the importance of considering multiple contexts *simultaneously*. For instance, there is growing interest in cross-classified multilevel modeling (CCMM) (Goldstein, 1994; Rabash and Browne, 2001), which allows researchers to examine instances when individuals are nested in non-hierarchical contexts, such as when students attending the same school live in different neighborhoods and conversely when students from the same neighborhood attend different schools. To date, CCMM has been used to examine the impact of schools and neighborhoods on a variety of health and behavioral outcomes (Dunn et al., 2015a, 2016, 2017, 2015b; Townsend et al., 2012; De Clercq et al., 2014; Evans et al., 2016), as

* Corresponding author at: Center for Applied Pediatric Quality Analytics, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA.
E-mail addresses: carly.milliren@childrens.harvard.edu (C.E. Milliren), cevans@uoregon.edu (C.R. Evans), tracy.richmond@childrens.harvard.edu (T.K. Richmond), edunn2@mgh.harvard.edu (E.C. Dunn).

https://doi.org/10.1016/j.healthplace.2018.05.009
Received 22 December 2017; Received in revised form 4 May 2018; Accepted 23 May 2018

well as contextual effects of classrooms and teachers on educational outcomes (Heck, 2009; Kim et al., 2010). A major advantage of CCMM relative to MLM is that it enables researchers to avoid the "omitted context bias", wherein variance in a random effects model is misattributed from the missing context to the included context, as the included context "soaks-up" the effect of the missing context (Dunn et al., 2015b; Evans et al., 2016).

Sample size requirements for MLM are well established (Dedrick et al., 2009; McNeish and Stapleton, 2014). To generate unbiased estimates of random effects variance parameters, methodologists recommend between 5 and 20 lower level units (e.g., students) as a minimum for each higher level unit (e.g., school) (McNeish and Stapleton, 2014). Including some schools with a smaller sample size in the data set is not problematic, however, because estimates for contexts with small sample size are automatically down-weighted in MLM estimation. Thus, *most* schools in the sample would need a minimum of 5–10 students to provide a reasonable estimate of the school-level variance. However, similar guidelines are not yet available for CCMM, raising questions about the minimum sample size required per unit of analysis in the CCMM setting.

Further, there is uncertainty about whether random effect estimates are sensitive to the sampling strategy and the potential imbalance of sample size across units of analysis. Many researchers conducting CCMM studies use data drawn from samples where only one context was originally intended to be studied. For instance, school-based researchers intentionally sample students by school, ignoring the distribution of students across neighborhoods. Because the dataset contained information about both school features and neighborhood of residence, researchers could fit a CCMM to estimate both school and neighborhood-level random effects – even though neighborhoods were not the primary sampling unit. As a result, the distribution of the sample across neighborhood catchment areas may be uneven due to schools being the primary sampling unit, potentially biasing estimates of neighborhood-level effects. Small and imbalanced neighborhood sizes could result in higher variability and imprecise estimates for random effects and possible bias leading to inaccurate conclusions regarding contextual effects. As CCMM becomes more popular with researchers encountering more non-nested data structures – particularly in the case of group randomized control trials – it is essential to determine whether estimates of contextual-level effects are biased when the sample sizes are unevenly distributed across the two contexts studied. If contextual effects are biased, it is also important to describe the direction of that bias, whether toward or away from the null.

The current study aimed to address these questions by performing a series of simulation analyses based on data from the National Longitudinal Study of Adolescent to Adult Health (Add Health) (Harris et al., 2009a, 2009b), one of the largest nationally representative surveys in the U.S (Harris et al., 2015). Our goal was to determine the extent to which a sample can be distributed unevenly across one higher-level context before random effects variance estimates become biased. Add Health was an ideal empirical dataset in which to ground these simulations because it is widely used in public health and has already linked contextual measures of schools and neighborhoods to health and behavior. Further, it intentionally sampled from one context (i.e., schools were the primary sampling unit) and the sample was distributed unevenly across a second context (i.e., neighborhoods). Additionally, because Add Health was drawn to be nationally representative, the distribution of students across schools and neighborhoods is a realistic sample of school catchment areas within the U.S. While schools in the sample each had a reasonable sample size, the neighborhoods those students came from were not always well represented, with many having only a single respondent. Furthermore, because of the rich information contextual information available, CCMMs are increasingly being used in Add Health papers despite unanswered questions of their validity prompted by the small neighborhood sample sizes. By anchoring these simulations to a realistic example and commonly used dataset, we ensure that our examination of CCMM validity is conducted within a relevant parameter space with practical implications for future Add Health studies. Body mass index (BMI) was chosen as the outcome for this simulation because of its clarity for analysis purposes (measured continuously and has an approximately Gaussian distribution) as well as its salience as a public health issue (Baskin et al., 2005; Lawrence, 2004).

2. Methods

Empirical data from the Wave 1 in-home sample of Add Health was used as a basis for the school and neighborhood data structure in our simulations. There were 20085 students who attended 132 unique schools and lived in 2410 unique neighborhoods. The school and neighborhood data structure in the Add Health is cross-classified because students attending the same school often resided in different neighborhoods and students living in the same neighborhood attended different schools. Specifically, there were 2979 unique combinations of school and neighborhood, with a median of 1 school per neighborhood (range 1–3) and a median of 14 neighborhoods per school (range 1–234). Thus, the data were not purely hierarchical, but rather schools in particular drew students from many neighborhoods.

Overall, school sizes in Add Health ranged from 20 to 1720 with median 126.5 (interquartile range 85–174.5). Neighborhood sizes ranged from 1 to 276 (median 2; interquartile range 1–5); 45% of neighborhoods had only a single student while only 8% had 25 or more. These values indicate a wide distribution in neighborhood sizes with most falling in the lower range. While Add Health schools would appear to have sufficient sample sizes, at least according to the rules for hierarchical MLM, it was unclear whether this highly imbalanced neighborhood design affects random effects variance estimates for neighborhoods in CCMM.

2.1. Assignment of students to neighborhoods for the simulation (determining balance)

To remain consistent with the existing cross-classified data structure, we maintained the number of students nested within each school (range 20–1720; mean 152; median 126), as well as the number of neighborhoods feeding into each school (range 1–234). With the structure defined, we sorted students into neighborhoods for three different levels of sample size balance across neighborhoods: *perfectly balanced*, *mildly imbalanced*, and *very imbalanced*.

For the perfectly balanced scenario, the number of students within each school was divided evenly across the neighborhoods sending students to that school. Due to rounding, some schools had too many or too few students; this was addressed by randomly subtracting or adding from neighborhoods so that the number of students in each school was consistent with the empirical data, each neighborhood still had at least one student, and as close to perfect balance as possible was achieved.

For both imbalanced scenarios, we utilized a geometric distribution to assign students to neighborhoods given the number of neighborhoods per school. The probability of assignment to a given neighborhood k given the initial proportion p, was calculated as:

$$P(X = k) = (1-p)^{k-1}p \quad (1)$$

where p = initial proportion (probability of assignment to first neighborhood) and k = given neighborhood sending students to a specific school. P was set at 0.25 for the mildly imbalanced and 0.7 for the imbalanced scenario, meaning that the first neighborhood for each school was assigned 25% of students and 70% of students, respectively. Fig. 1 illustrates the assignment of students to neighborhoods under the balance scenarios for a hypothetical school with 60 students from 12 neighborhoods.

In practice, under both the mildly imbalanced and very imbalanced scenarios this resulted in some neighborhoods with zero students