Appendix 1

A discussion of model selection

In our database, children are nested within sites, raising the issues of how best to address this clustering. We considered several options, including Generalized Estimating Equations (GEE); multi-level models (MLMs); random-effects (RE) models; and fixed-effects (FE) models. Each of these models has certain advantages (Wooldridge 2001).

GEE is an unconditional method (the effects of individual-level covariates are estimated without regard to which cluster they belong) so the estimates are interpreted as “population-averaged effects.” GEE is popular due to the ease of estimation (e.g., the widely available “cluster robust” option in Stata is a special case of GEE estimation) and because it does not require the distributional assumptions that most MLMs do.

MLMs are conditional methods in which the analysis essentially conditions on the cluster in order to estimate the effects of individual-level factors, averaging the effects across clusters. MLM’s explicitly try to model the variance that occurs at different levels of the hierarchy, so they can provide information on what percent of the explained variance is accounted for at each level. MLMs are more efficient than GEE if the distributional assumptions hold. Due to their “shrinkage” estimation properties, MLMs are also preferred when the goal of the analysis is to profile or rank clusters.

RE models are a special case of MLMs in which the covariance between the error terms at the two levels is assumed to be zero; the slope coefficients are fixed; and the intercept term is allowed to be random but not a function of higher-level covariates. RE models are often used to address clustering when a full-fledged MLM is not warranted, the observations can be assumed to be random draws from a large population, and the (unobserved) cluster-level heterogeneity can be assumed to be uncorrelated with the covariates. Relative to FE models, RE models are more efficient because the FE estimator uses only the “within-group” variation, but the RE estimator is based on an optimal combination of the “within-group” and “between-group” estimators.

In FE models, each cluster has its own (fixed) intercept term. In linear models, this is equivalent to “de-meaning” both the dependent and independent variables, so that the estimator relies solely on the “within-group” variation. Note that as groups become larger, the information from “within-group” variance increases, becoming relatively more important than the “between-group” variance, and the FE and RE estimators eventually converge.

Our choice of the FE models over the other alternatives was based on several considerations. Our paper focuses on the main effects of race/ethnicity, rather than on variation in race/ethnicity effects by site (moderation) or profiling/ranking of sites. Sample sizes by race/ethnicity were somewhat limited at the site level, so potential overfitting was a concern. Most importantly, we considered the biggest threat to the validity of our findings to be potential confounding of the child’s race with unobserved site-level heterogeneity, thereby violating one of the assumptions of the MLM and RE models (Ebbes 2004; Hanchane and Mostafa, 2010; Greene 2014). For example, it seems likely that minority children might be more likely to live in areas with (unobservably) more limited provider supply or worse quality of care. In this case, FE models would still yield unbiased estimates of the effects of race/ethnicity (because the estimation uses only the variation across children of different race/ethnicity located within the same site and does not compare children of different race/ethnicity located in different sites). In contrast, the other models might not, since part of the race/ethnicity effect is estimated using the between-site variation, i.e., the race/ethnicity effect would pick up not just differences in the experience of minority and non-minority children in a given location, but might also pick up the effect of being in a location with worse access and quality. Other authors have demonstrated the advantage of FE models when heterogeneity at the higher level cannot be assumed to be uncorrelated with the lower-level predictor of interest (Hanchane and Mostafa, 2010; Bao, Fox and Escarce, 2007; Boone-Heinonen et al., 2011). Given that our primary concern was to obtain an unbiased estimate of the (main) effect of race/ethnicity, FE models appeared to be the most conservative choice, providing greater confidence that our estimates are not confounded by unmeasured site characteristics at the expense of less efficiency in estimation and less ability to generalize to the entire underlying population.

References

Bao Y, Fox SA, Escarce JJ. Socioeconomic and racial/ethnic differences in the discussion of cancer screening: "between-" versus "within-" physician differences. Health Serv Res 2007 Jun;42(3 Pt 1):950-70.

Boone-Heinonen J, Diez Roux AV, Kiefe CI, Lewis CE, Guilkey DK, Gordon-Larsen P. Neighborhood socioeconomic status predictors of physical activity through young to middle adulthood: the CARDIA study. Soc Sci Med. 2011;72(5):641-9.

Ebbes, P., Böckenholt, U., & Wedel, M. (2004). Regressor and random effects dependencies in multilevel models. Statistica Neerlandica, 58(2), 161-178.

Greene W. Fixed and random effects in nonlinear models. In: Economics Working Papers. Accessed July 2014 at **https://archive.nyu.edu/bitstream/2451/26204/2/1-1.pdf**

Hanchane, S and Mostafa, T (2010). *Endogeneity Problems in Multilevel Estimation of Education Production Functions: an Analysis Using PISA Data* published by the Centre for Learning and Life Chances in Knowledge Economies and Societies at: http://www.llakes.org

Wooldridge J. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press; 2002.