Understanding, Choosing, and Unifying Multilevel and Fixed Effect Approaches

Chad Hazlett[1] and Leonard Wainstein[2]

[1] *Assistant Professor, Departments of Statistics and Political Science, University of California Los Angeles, CA, USA. Email: chazlett@ucla.edu, URL: http://www.chadhazlett.com*
[2] *PhD Candidate, Department of Statistics, University of California Los Angeles, CA, USA. Email: lwainstein@ucla.edu*

Abstract

When working with grouped data, investigators may choose between "fixed effects" models (FE) with specialized (e.g., cluster-robust) standard errors, or "multilevel models" (MLMs) employing "random effects." We review the claims given in published works regarding this choice, then clarify how these approaches work and compare by showing that: (i) random effects employed in MLMs are simply "regularized" fixed effects; (ii) unmodified MLMs are consequently susceptible to bias—but there is a longstanding remedy; and (iii) the "default" MLM standard errors rely on narrow assumptions that can lead to undercoverage in many settings. Our review of over 100 papers using MLM in political science, education, and sociology show that these "known" concerns have been widely ignored in practice. We describe how to debias MLM's coefficient estimates, and provide an option to more flexibly estimate their standard errors. Most illuminating, once MLMs are adjusted in these two ways *the point estimate and standard error for the target coefficient are exactly equal to those of the analogous FE model with cluster-robust standard errors.* For investigators working with observational data and who are interested only in inference on the target coefficient, either approach is equally appropriate and preferable to uncorrected MLM.

Keywords: multilevel models, hierarchical models, fixed effects, random effects, grouped data, cluster-robust standard errors

1 Introduction

Researchers in many applied fields encounter data structures with observations that are grouped or clustered in one or more ways, for example students nested in classrooms and/or schools, and perhaps measured repeatedly over time. Such observations are referred to variably as grouped, clustered, hierarchical, multilevel, or repeated measures, and include panel, longitudinal, or time-series cross-sectional data. In one very common context—and the primary example considered here—investigators hope to estimate the effect of some "treatment," or target covariate, that varies at a lower level while accounting for confounding that is hoped to be fixed at the higher level. For example, in country–year data, some countries may adopt a treatment in some years, and we seek to account at least for country- (group-) level confounders, in the hopes that remaining, time-varying confounding is absent or less problematic.

Multilevel data structures, however, pose complications for estimation and inference by violating the independence of observations assumed under classical inference. Although a longstanding and ubiquitous issue, methodological practices for dealing with this non-independence have not converged across disciplinary traditions. One tradition, often referred to as the "fixed effects" approach (FE), advises investigators to account for group-level confounders by introducing group-level, freely varying, intercepts to their models. The nonindependence of observations then complicates only variance estimation, so investigators are instructed to choose a variance estimator that accommodates the forms of intragroup dependency assumed to exist. One popular choice, closely examined here, is the "cluster-robust standard error" (White 1984).

Political Analysis (2022)
vol. 30: 46–65
DOI: 10.1017/pan.2020.41

Published
31 December 2020

Corresponding author
Chad Hazlett

Edited by
Jeff Gill

© The Author(s) 2020. Published by Cambridge University Press on behalf of the Society for Political Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

A different tradition holds that multilevel data must be analyzed with “multilevel models” (MLMs). These models look similar to FE models, but include terms that are estimated as “random effects,” meaning that the coefficients are assumed to have a distribution rather than to be fixed in truth. In a review of 10 textbooks and 4 well-cited pedagogical articles[1] as well as 109 empirical articles employing MLMs in top education, political science, and sociology journals,[2] we find four common reasons given to employ MLM rather than FE: (1) MLMs correctly estimate standard errors in grouped data; (2) MLMs estimate coefficients more efficiently and produce more accurate predictions than do the analogous FE models; (3) MLMs allow intercepts and slopes to vary by group, like analogous FE models; but (4) unlike those FE models, MLMs can estimate coefficients for group-level variables (and their interactions with lower-level variables) while allowing varying intercepts (and slopes). By contrast, particularly for fields closely aligned with econometrics, both long-standing (e.g., Hausman 1978) and more recent work (e.g., Clark and Linzer 2015) emphasize bias concerns with MLMs. In the common setting where one particular variable is the treatment, MLM estimates can have lower variance than do FE estimates for the coefficient of interest, but are biased when group-level confounding is present, compelling users to employ FE.[3]

The choice between these approaches remains a matter of debate and disciplinary tradition, and is sometimes justified based on erroneous claims as to these models’ properties. In this paper, we seek to demystify their differences, the problems associated with them, and ultimately important equivalences between them. First, to illuminate a key difference between the approaches, we show that random effect estimates in MLMs are precisely equivalent to FE estimates that have been shrunken through a *regularization* process, that is, by penalizing larger coefficients. This explains the principle concern with random effects: bias that emerges because these variables are not “allowed” to adjust for confounding as intended. Because regularization reduces over-fitting, this also demystifies why MLM’s out-of-sample outcome predictions are typically more accurate. Second, this bias can be eliminated in many cases by a long-standing adjustment from Mundlak 1978. For models with group-level intercepts added as random effects (“random intercept” models), which we focus on here, adding the group-level averages of all included variables as regressors (or a variety of equivalent procedures such as group-wise centering) relieves the bias otherwise induced by this regularization. We also present a more general solution for more complex models. Finally, we consider variance estimation. Contrary to claims we document, MLMs do not automatically ensure appropriate standard error estimates for grouped data. Rather, the standard MLM approach relies on strict assumptions and can have poor coverage in even simple circumstances.

We are not the first to draw many of these conclusions. However, our review of empirical papers employing MLMs shows widespread failure to either appreciate the concerns they raise or employ suggested solutions. Bringing these concerns together, we describe a “bias-corrected MLM” that employs cluster-robust standard errors, which make no assumption about within-group dependency and assume no between-group dependence. Remarkably, once these adjustments are

[1] The textbooks are: Faraway (2016), Finch, Bolin, and Kelley (2016), Fitzmaurice, Laird, and Ware (2004), Greene (2003), Heck, Thomas, and Tabata (2013), Luke (2004), Snijders and Bosker (2011), Hox and Roberts (2011), Gelman and Hill (2006), Hoffman (2015). The articles are: Snijders and Berkhof (2008), Raudenbush (2009), Gelman (2006), Steenbergen and Jones (2002).

[2] American Journal of Political Science (17 articles), the American Political Science Review (13), the Journal of Politics (20), the American Education Research Journal (28), Educational Evaluation and Policy Analysis (8), the American Journal of Sociology (13), and the American Sociological Review (10). We decided on this selection of journals after asking specialists in each field about the top journals that often publish papers that employ MLM. To find the articles, we searched on “multilevel,” “multi-level,” “hierarchical,” “random effect,” “random effects,” “random-effect,” and “random-effects.” Our political science and sociology reviews currently cover all articles dated January 2017 through December 2018, and our education review currently covers all articles dated January 2017 through April 2019.

[3] When this bias is present, the root mean square error of MLM may be higher than that of FE even if MLM has lower variance, as we demonstrate. It is also possible to construct cases wherein a MLM and FE model have equal variance. One example is when treatment assignment is “perfectly blocked” within group, so that the estimated covariance of the treatment and block indicators is numerically zero in every sample. In this case, FE, MLM with random intercepts, and even OLS produce identical estimates and thus have identical efficiency. We thank Ian Lundberg for pointing this out.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

made, the resulting MLM produces coefficient and standard error estimates *identical* to those of the analogous FE model with cluster-robust standard errors. We regard this as the most important conclusion, since for investigators interested only in the target coefficient and its standard error, it resolves any debate as to which approach is more appropriate. In most cases, we thus recommend either of these unbiased approaches over uncorrected MLM, due to the dangers of heightened bias and root mean square error when MLM's strict assumptions are violated. The remaining differences between these approaches are that MLM has better (out-of-sample) accuracy for the outcome predictions, and allows the user to estimate coefficients that would otherwise be dropped by FE (e.g., group-level covariates), although interpreting these coefficients can be problematic.

2 Background

2.1 Notation

We first set notation. To aid the reader, Appendix A.1 describe (i) symbols used in our notation (Table 1) and (ii) abbreviations we use relating to models (Table 2).

Let $g = 1, \ldots, G$ index the group. We index vectors belonging to group g with the subscript g, for example, the outcome for all units in group g is given by the vector Y_g. The ith unit in group g is then indexed by $g[i]$, for example, unit i in group g has outcome $Y_{g[i]}$. This notation reminds the reader that unit i is contained in group g. Each group g has size n_g with $N = \sum_{g=1}^{G} n_g$. Let $X_{g[i]}$ be the p-dimensional vector of covariates, including an intercept term, with an associated coefficient vector β. One element of $X_{g[i]}$ in particular will be regarded as a treatment in settings described here. X_g is the matrix of $X_{g[i]}$ for group g, and X is the matrix of $X_{g[i]}$ for the entire sample.

$$X_{g[i]} = \begin{bmatrix} 1 \\ X_{g[i]}^{(1)} \\ \vdots \\ X_{g[i]}^{(p-1)} \end{bmatrix} \in \mathbb{R}^{p}, \; X_g = \begin{bmatrix} X_{g[1]}^{\top} \\ \vdots \\ X_{g[n_g]}^{\top} \end{bmatrix} \in \mathbb{R}^{n_g \times p}, \; X = \begin{bmatrix} X_1 \\ \vdots \\ X_G \end{bmatrix} \in \mathbb{R}^{N \times p}, \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \in \mathbb{R}^{p}$$

Similarly, $Z_{g[i]}$ is a d-dimensional vector of covariates, often containing a subset of the covariates in $X_{g[i]}$, and possibly an intercept which will later function as an indicator of membership to group g. For each group g, the $Z_{g[i]}$ have an associated coefficient vector γ_g. Z_g is the matrix of $Z_{g[i]}$ for group g, Z is a block diagonal matrix of the Z_g, and γ stacks the set of γ_g into a matrix.

$$Z_{g[i]} = \begin{bmatrix} Z_{g[i]}^{(0)} \\ \vdots \\ Z_{g[i]}^{(d-1)} \end{bmatrix} \in \mathbb{R}^{d}, \; Z_g = \begin{bmatrix} Z_{g[1]}^{\top} \\ \vdots \\ Z_{g[n_g]}^{\top} \end{bmatrix} \in \mathbb{R}^{n_g \times d}, \; Z = \begin{bmatrix} Z_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & Z_G \end{bmatrix} \in \mathbb{R}^{N \times Gd},$$

$$\gamma_g = \begin{bmatrix} \gamma_{0g} \\ \vdots \\ \gamma_{(d-1)g} \end{bmatrix} \in \mathbb{R}^{d}, \; \gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_G \end{bmatrix} \in \mathbb{R}^{Gd}$$

As noted, $Y_{g[i]}$ is the outcome of interest, and $\epsilon_{g[i]}$ is its associated residual/error term. Y and ϵ are $N \times 1$ vectors containing $Y_{g[i]}$ and $\epsilon_{g[i]}$ for the entire sample.

$$Y_{g[i]} \in \mathbb{R}, \; Y_g = \begin{bmatrix} Y_{g[1]} \\ \vdots \\ Y_{g[n_g]} \end{bmatrix} \in \mathbb{R}^{n_g}, \; Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_G \end{bmatrix} \in \mathbb{R}^{N}, \; \epsilon_{g[i]} \in \mathbb{R}, \; \epsilon_g = \begin{bmatrix} \epsilon_{g[1]} \\ \vdots \\ \epsilon_{g[n_g]} \end{bmatrix} \in \mathbb{R}^{n_g}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_G \end{bmatrix} \in \mathbb{R}^{N}$$

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

2.2 Fixed effect and multilevel models

Varying intercepts: Group fixed effects and random intercept models. We focus mainly on uses of FE and MLM that allow each group in the data a different intercept but no other group-varying coefficients.[4] Suppose data are generated according to the simple model, written in matrix form,

$$Y = X\beta + Z\gamma + \epsilon, \tag{1}$$

where Z contains only indicators for membership to each group g, that is, each Z_g is a vector of ones, meaning that only intercepts may differ between groups. While useful to write in this form because it shows the role of Z, the model is often expressed as

$$Y_{g[i]} = X_{g[i]}^{\top}\beta + \gamma_g + \epsilon_{g[i]}, \tag{2}$$

where γ_g are group-specific deviations from the overall intercept in β. Although a model can be fitted regardless of the correlation between the residuals and $X_{g[i]}$, investigators are usually interested in understanding the effect of a treatment variable (in $X_{g[i]}$) on $Y_{g[i]}$, which requires the "conditional independence assumption," $\mathbb{E}(\epsilon_{g[i]} \mid X, Z) = 0$. This is an identification assumption relating to the form of confounding, although it is also subject to model specification. We describe the types of permitted confounding that are used to justify this assumption in Section 2.3.

The distinction between FE and MLM for the varying intercepts model in Equation (2) relates to the assumptions they employ during estimation. Both FE and MLM regard β as "fixed," that is, nonrandom with no distributional assumption. However, FE and MLM diverge in their handling of γ_g. In FE, the γ_g are regarded as fixed, like β. FE thus estimates β and γ by including indicator variables for group membership (in Z) as additional regressors in an OLS of Y on X, or by equivalent demeaning/partialing-out procedures. We refer to this as "group fixed effects" (**Group-FE**) here. In practice, one group indicator and its corresponding γ_g must be dropped, or the intercept term must be dropped from $X_{g[i]}$.

MLM, by contrast, treats the γ_g as "random effects," meaning that each γ_g is estimated under the assumption that it may have a distribution, that is has nonzero variance rather than being a fixed quantity. This has implications for estimating both $\gamma and \beta$, as well as for constructing variance estimates, detailed in Section 2.4.[5] Specifically, we define the "random intercept" (**RI**) model,

$$Y_{g[i]} = X_{g[i]}^{\top}\beta + \gamma_g + \epsilon_{g[i]}, \quad \gamma_g \mid X, Z \overset{iid}{\sim} N(0, \omega^2) \tag{RI}$$

where we also assume that $\epsilon_{g[i]} \perp\!\!\!\perp \gamma_{g'} \mid X, Z$ for all g, g', and i. Additionally, the conditional independence assumption is elaborated to require a multivariate-normal distribution for the residual vectors in each group, $\epsilon_g \mid X, Z \overset{ind}{\sim} N(0, \Sigma_g)$ where $\Sigma_g \in \mathbb{R}^{n_g \times n_g}$ is group g's error covariance matrix. All γ_g can be kept and estimated by this approach, together with an intercept in β. It is also commonly assumed that the $\epsilon_{g[i]}$ are spherical (i.e., $\Sigma_g = \sigma^2 I_{n_g}$), although we avoid that restriction unless otherwise noted.

[4] This is perhaps the most common use of MLM: In our review of empirical papers employing MLM in education, political science, and sociology journals, MLM models solely allowing group-varying intercepts ("random intercept" models, described momentarily) were by far the most common usage of MLM, covering at least 24 of 36 articles in education, at least 39 of 50 in political science, and at least 21 of 23 in sociology.

[5] While the terms "fixed" and "random" as defined above are in keeping with common usage in the MLM literature, their meanings can vary widely and sometimes even conflict (see Gelman *et al.* 2005 for examples). Random effect intercepts (or later, coefficients) are also sometimes referred to as "modeled" (e.g., Gelman 2006) because they are given a probability model (which can also be modified, e.g., centered on a function of covariates rather than zero).

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

General fixed effect and multilevel models. Although we focus mainly on Group-FE and RI models here, the claims made below pertain to FE and MLM in their more general form,

$$Y_{g[i]} = X_{g[i]}^{\top}\beta + Z_{g[i]}^{\top}\gamma_g + \epsilon_{g[i]}, \tag{3}$$

where the γ_g represent group-specific coefficients for the variables in Z_g, which may now include variables other than group indicators as above. As before, FE estimates both β and γ by OLS regression of Y on X and Z. For identifiability, the covariates that $X_{g[i]}$ and $Z_{g[i]}$ share (or that result in perfect colinearity) are either dropped from $Z_{g[i]}$ for one g, or dropped from $X_{g[i]}$.

When instead fit with random effects in MLM, the coefficients γ_g are estimated with additional distributional assumptions, specifically

$$Y_{g[i]} = X_{g[i]}^{\top}\beta + Z_{g[i]}^{\top}\gamma_g + \epsilon_{g[i]}, \quad \gamma_g \mid X, Z \overset{iid}{\sim} \mathcal{N}(0, \Omega), \tag{MLM}$$

where $\Omega \in \mathbb{R}^{d\times d}$ and we assume $\epsilon_{g[i]} \perp\!\!\!\perp \gamma_{g'} \mid X, Z$ for all g, g', and i.[6] In addition, we define $\text{var}(\epsilon \mid X, Z) = \Sigma$ to be the ordered block diagonal matrix of Σ_g, and $\Omega_{\text{block}} = \text{var}(\gamma \mid X, Z)$ to be the block diagonal matrix of Ω,

$$\Sigma = \begin{bmatrix} \Sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Sigma_G \end{bmatrix} \in \mathbb{R}^{N\times N}, \; \Omega_{\text{block}} = \begin{bmatrix} \Omega & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Omega \end{bmatrix} \in \mathbb{R}^{Gd\times Gd}$$

2.3 Identification: From nonparametric conditions to specification requirements

Let us assume there is no within-group confounding. In longitudinal settings, this is to say there are no unobserved time-varying confounders. Such an assumption ensures non-parametric identifiability, that is, the causal effect of the treatment can be identified conditionally on group and then averaged together as desired across groups.[7]

For such nonparametric identification to be sufficient for unbiased estimation, however, would require the ability to condition on group non-parametrically (i.e., estimate the relationship between treatment and outcome within each group separately). In most settings the investigator is unable to do this, and so turns to modeling assumptions that instead “account for” group in a specific model. Because of this model dependence, identification of treatment effects then requires additional assumptions related to the specification of those models.

The traditional motive for Group-FE is an assumption that the only source of confounding is group-level confounding that takes the linear form of Equation (2): a constant, additive shift by group (i.e., the γ_g). The conditional independence assumption needed is $\mathbb{E}(\epsilon_{g[i]} \mid X, Z) = 0$ (e.g., Greene 2003). The Group-FE approach is powerful precisely because including Z fully purges the group-level intercepts, γ_g, from this residual, thus removing confounding. So long as other (within-group) forms of confounding do not exist, this unbiasedly estimates β.

The central concern with RI is that *even under the conditional independence assumption, random effect estimation fails to remove this confounding*. As we explain in Section 3.1, this is because RI does not fully account for the γ_g, biasing the estimate of β as well. Avoiding this bias requires *additionally* that the “uncorrelated random effects” assumption be true: that the γ_g are

[6] Unless otherwise noted, we assume the γ_g are identically distributed with common variance Ω, although this can be relaxed (Section 3.3).

[7] In terms of directed acyclic graphs (DAGs; Pearl 2000), conditioning on group must block all backdoor paths between the treatment and the outcome. In terms of potential outcomes, the potential outcomes at different observations within groups (e.g., years within country) must be independent of treatment conditional on group. See Imai and Kim (2019) for a recent discussion of uncommonly recognized ways that the no-unobserved-confounding assumption may be violated.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

uncorrelated with $X_{g[i]}$, so that failure to account for γ_g does not bias estimates of β.[8,9] This is problematic, since γ_g is most important to account for when it is correlated with the treatment variable of interest in $X_{g[i]}$ and thus acts as a confounder. While this assumption, also referred to simply as the "random effects assumption" (Bell and Jones 2015; Kim and Steiner 2019), is well-known in principle, it remains widely neglected in practice (see Section 3.1) despite attractive solutions (see Section 3.2).

2.4 Parameter estimation in MLM

We begin by briefly reviewing how MLM parameters are estimated.[10] Estimation proceeds in three steps: (1) estimate Ω and Σ, (2) estimate β using the estimate of (Ω, Σ), and (3) estimate γ using the estimate of (β, Ω, Σ). The first two steps require the distribution of Y given X and Z. Because, in MLM, the $Z_{g[i]}^{\top}\gamma_g$ are mean-zero given X and Z, the $(Z_{g[i]}^{\top}\gamma_g + \epsilon_{g[i]})$ can be treated as combined mean-zero error terms. This allows one to formulate MLM on the sample-level as $Y = X\beta + \epsilon^*$ where $\epsilon^* = Z\gamma + \epsilon$. Then, because γ and ϵ are both normally distributed given X and Z, we have

$$Y \mid X, Z \sim \mathcal{N}(X\beta, V) \quad \text{where} \quad V = \text{var}(\epsilon^* \mid X, Z) = Z\Omega_{\text{block}}Z^{\top} + \Sigma. \tag{4}$$

The likelihood is then

$$L(\beta, \Omega, \Sigma \mid Y, X, Z) = p(Y \mid X, Z, \beta, \Omega, \Sigma) \propto |V|^{-1/2}\exp\left(-\frac{1}{2}(Y - X\beta)^{\top}V^{-1}(Y - X\beta)\right). \tag{5}$$

Given a choice of (Ω, Σ), which determines V, maximizing the likelihood for β would yield

$$\hat{\beta}(\Omega, \Sigma) = (X^{\top}V^{-1}X)^{-1}X^{\top}V^{-1}Y. \tag{6}$$

However, estimates of Ω and Σ do not enjoy similarly simple closed solutions. Instead, they are typically found iteratively through either unrestricted maximum likelihood estimation or restricted maximum likelihood estimation, both using Equation (5). β can then be estimated by plugging estimates of Σ and Ω into Equation (6). Finally, γ is estimated by maximizing its posterior probability given Y, X, Z, and the estimated (β, Ω, Σ), that is,

$$\hat{\gamma}(\beta, \Omega, \Sigma) = \arg\max_{\gamma} p(\gamma \mid Y, X, Z, \beta, \Omega, \Sigma) = \Omega_{\text{block}}Z^{\top}V^{-1}(Y - X\beta). \tag{7}$$

In this article, we focus solely on estimation and inference for β and γ. Because MLM's estimates of these parameters are functions of Ω and Σ, our results hold regardless of the choice between unrestricted or restricted maximum likelihood in estimating Ω and Σ. For this reason, we refer to arbitrary MLM estimates of the parameters as $\hat{\Omega}_{\text{MLM}}$, $\hat{\Sigma}_{\text{MLM}}$, $\hat{\beta}_{\text{MLM}}$, and $\hat{\gamma}_{\text{MLM}}$.

Finally, treating $\hat{V}_{\text{MLM}}$ as fixed, the conditional variance of $\hat{\beta}_{\text{MLM}}$ is simply

$$\text{var}(\hat{\beta}_{\text{MLM}} \mid X, Z) = (X^{\top}\hat{V}_{\text{MLM}}^{-1}X)^{-1}X^{\top}\hat{V}_{\text{MLM}}^{-1}\text{var}(Y \mid X, Z)\hat{V}_{\text{MLM}}^{-1}X(X^{\top}\hat{V}_{\text{MLM}}^{-1}X)^{-1}, \tag{8}$$

[8] The presumption that γ_g and $X_{g[i]}$ are uncorrelated is sometimes described less as an "assumption" and instead as a feature of a "working model" or a prior belief, that is, a convenience that we employ but do not necessarily expect true, or that becomes irrelevant as the sample size grows. However, as shown here, it is a consequential modeling decision and, in finite samples, demonstrably leads to undesirable estimates.

[9] In the more general setting where the entire "random effect contribution" is captured by $Z_{g[i]}^{\top}\gamma_g$ (i.e., Equation 3), this implies that $Z_{g[i]}^{\top}\gamma_g$ is uncorrelated with $X_{g[i]}$.

[10] Both frequentist (maximum likelihoood) and Bayesian estimation approaches are available. The former is the most commonly taught and employed (e.g., `lme4` in `R`), at least for the simpler models we consider here. An excellent review of Bayesian MLM estimation can be found in Gelman and Hill (2006).

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

which is commonly simplified to $(X^\top \hat{V}_{\text{MLM}}^{-1} X)^{-1}$ as the standard model-based MLM variance estimator, by assuming that $\hat{V}_{\text{MLM}}$ is a consistent estimator of $\text{var}(Y \mid X, Z)$.

3 Analytical insights

We now analyze three features of MLM that aid in understanding how these methods compare.

3.1 Random effects as regularization and the "incomplete conditioning" problem

The first piece shows an equivalence between the random effects employed in MLM and regularization, that is, a penalized fitting approach. We begin by introducing the "regularized fixed effects" (**regFE**) class of models. Suppose we are interested in optimal out-of-sample generalization. Then, instead of estimating Equation (2) by minimizing (in-sample) squared error, we minimize the squared error plus a penalty proportional to the sum of squared γ_g values, known as L_2 or Tikhonov regularization:

$$(\hat{\beta}_{\text{regFE}}, \hat{\gamma}_{\text{regFE}}) = \underset{\beta,\gamma}{\arg\min} \Big(\sum_{g=1}^{G} \sum_{i=1}^{n_g} [Y_{g[i]} - X_{g[i]}^\top \beta - \gamma_g]^2 + \lambda \sum_{g=1}^{G} \gamma_g^2 \Big), \tag{9}$$

where $\lambda > 0$ determines the extent of regularization on γ_g—larger values shrink each γ_g toward to 0, while smaller values yield estimated intercepts closer to Group-FE's estimates—and may be chosen by various means such as some form of cross-validation.[11] The key result is that for a particular choice of λ, such a regularized model is equivalent to the RI model,

Theorem 3.1 (Equivalence of RI and regFE) Let $(\hat{\omega}_{\text{RI}}^2, \hat{\Sigma}_{\text{RI}}, \hat{\beta}_{\text{RI}}, \hat{\gamma}_{\text{RI}})$ be estimates from the RI model. If (i) $\Sigma = \sigma^2 I_N$ and (ii) $\lambda = \hat{\sigma}_{\text{RI}}^2 / \hat{\omega}_{\text{RI}}^2$ in Equation (9), then $(\hat{\beta}_{\text{regFE}}, \hat{\gamma}_{\text{regFE}}) = (\hat{\beta}_{\text{RI}}, \hat{\gamma}_{\text{RI}})$.

We omit the proof here, as Theorem 3.2 below offers a generalization. That MLM provides "shrinkage" or "partial-pooling" estimates is very well-known (e.g., Steenbergen and Jones 2002; Fitzmaurice, Laird, and Ware 2004; Luke 2004; Gelman 2006), and in recent texts MLM has even been referred to as employing a "regularizing prior" (e.g., McElreath 2018). The equivalence between regularization and holding a prior on the regularized coefficients is also well known.[12] That said, because we did not find any explicit formalization of the equivalence of estimation in MLM to regularization with a specific choice of λ in any textbooks or articles we reviewed, we fill this gap.

We demonstrate with a simple simulation example, drawing 1,000 datasets from the following data generating process (DGP) with $G = 25$ and $n_g = 15$ each time:

$$Y_{g[i]} = \beta_0 + \beta_1 X_{g[i]} + \gamma_g + \epsilon_{g[i]} \quad \text{where} \quad X_{g[i]} \overset{iid}{\sim} N(0,1)\,,\ \gamma_g \overset{iid}{\sim} N(0,1), \tag{10}$$

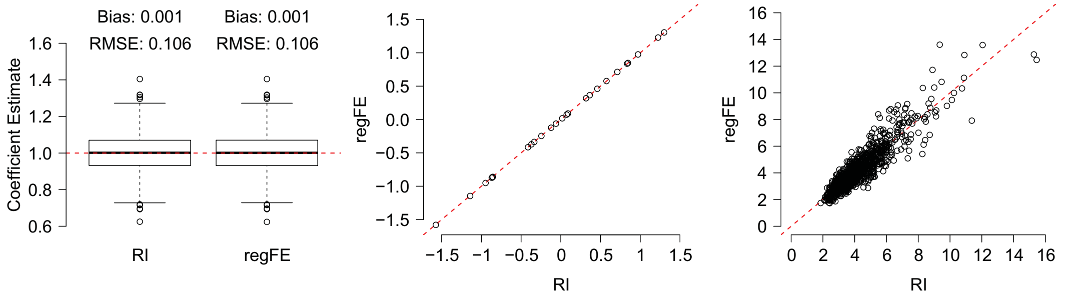
where λ for regFE is chosen by cross-validation and $(\hat{\sigma}_{\text{RI}}^2, \hat{\omega}_{\text{RI}}^2)$ for RI by restricted maximum likelihood.[13] We see in Figure 1 that the estimates of β and γ are nearly identical. They would be

[11] Cross-validation partitions the data into "folds," using all of the folds except for one to estimate a model, and the held-out fold to make predictions with said model. This process is repeated, holding out each fold in turn, ultimately producing an estimate for every observation from a model that was not trained on that observation. These predictions can be used to approximate out-of-sample error. Here we use ten-fold cross-validation to choose the λ that minimizes this out-of-sample error estimate.

[12] Specifically, regularization with the L_2 penalty as used here is equivalent to finding the maximum *a posteriori* (MAP) estimates under a prior that $\gamma_g \mid X, Z \overset{iid}{\sim} N(0, \sigma^2/\lambda)$. We use this connection in the proof of the equivalence between MLM and regFE. Note that other regularizing norms could be used. For example, regularization with an L_1 norm, $\sum_g |\gamma_g|$, produces the MAP estimator when we hold a Laplacian prior on γ_g, and would have the effect of inducing sparsity, possibly shrinking some γ_g to exactly zero.

[13] Here, and elsewhere where it is clear from context, we refer to the variable of interest, $X_{g[i]}^{(1)}$, as simply $X_{g[i]}$ to reduce notation, neglecting that $X_{g[i]}$ may also include an intercept or other terms.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press



(a) β (b) γ (c) λ

Figure 1. (a) Boxplots of $\hat\beta_1$ from RI and regFE over 1,000 iterations. The dashed line indicates the true value for β_1 (i.e., $\beta_1 = 1$). The estimates from each method are correlated at 0.999. (b) Example for the estimated intercepts from the RI model ($\hat\gamma_{\text{RI}}$) or from regFE ($\hat\gamma_{\text{regFE}}$) in a single iteration of the simulation. The points correspond to the coordinates $(\hat\gamma_{g,\text{RI}}, \hat\gamma_{g,\text{regFE}})$ across g, and the dashed line indicates equality. The RI and regFE estimates have a correlation of 0.999. (c) Plot of λ from RI and regFE over 1,000 iterations. At each iteration, λ is chosen by cross-validation for regFE, and $\lambda = \hat\sigma^2_{\text{RI}}/\hat\omega^2_{\text{RI}}$ for RI. The dashed line indicates equality. The λ from RI and regFE have means of 4.278 and 4.277, respectively, and are correlated at 0.910.

numerically equal if, instead of setting λ in regFE by cross-validation, we chose $\lambda = \hat\sigma^2_{\text{RI}}/\hat\omega^2_{\text{RI}}$ as per Theorem 3.1. However, despite the different selection procedures, we also see from Figure 1 that the λ from RI and regFE are quite similar. This is because while $\lambda = \hat\sigma^2_{\text{RI}}/\hat\omega^2_{\text{RI}}$ in RI is not made specifically with predictive accuracy in mind, the choice is a sensible one from a prediction point of view. To see this, note first that the portion of $Y_{g[i]}$ that is not explained by $X_{g[i]}$ is the sum of γ_g and $\epsilon_{g[i]}$, which have variances of ω^2 and σ^2, respectively. When, for example, the γ_g have high variance in relation to that of $\epsilon_{g[i]}$, $\lambda = \sigma^2/\omega^2$ will be small, which is appropriate since γ_g will be helpful in predicting $Y_{g[i]}$. The converse is true when the γ_g have low variance relative to $\epsilon_{g[i]}$, leading to a higher effective λ and desirable shrinkage on γ_g to avoid over-fitting.

More generally, consider estimating Equation (3) with the regularized regression:

$$(\hat\beta_{\text{regFE}}, \hat\gamma_{\text{regFE}}) = \arg\min_{\beta,\gamma}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}[Y_{g[i]} - X_{g[i]}^\top\beta - Z_{g[i]}^\top\gamma_g]^2 + \sum_{g=1}^{G}\gamma_g^\top\Lambda\gamma_g\right), \tag{11}$$

where $\Lambda \in \mathbb{R}^{d\times d}$ is symetric and positive semi-definite.

Again, Λ may be obtained by various means, such as cross-validation.[14] The equivalence with MLM is then given in Theorem 3.2.

Theorem 3.2 (General equivalence of MLM and regFE) Let $(\hat\Omega_{\text{MLM}}, \hat\Sigma_{\text{MLM}}, \hat\beta_{\text{MLM}}, \hat\gamma_{\text{MLM}})$ be the estimates from MLM. If (i) $\Sigma = \sigma^2 I_N$ and (ii) $\Lambda = \hat\sigma^2_{\text{MLM}}\hat\Omega^{-1}_{\text{MLM}}$ in Equation (11), then $(\hat\beta_{\text{regFE}}, \hat\gamma_{\text{regFE}}) = (\hat\beta_{\text{MLM}}, \hat\gamma_{\text{MLM}})$.

Proof of Theorem 3.2 is given in Appendix A.3 and proves Theorem 3.1 as a special case with $Z_{g[i]} = [1]$. This equivalence is useful in comprehending several of MLM's most central advantages, and limitations. First, the equivalence to regularization immediately explains why

[14] Note that the procedure in Equation (11) penalizes the magnitude of γ_g, as $\gamma_g^\top\Lambda\gamma_g \geq 0$ because Λ is positive semi-definite. Note also that if $Z_{g[i]} = [1]$, then Equations (9) and (11) coincide — γ_g and Λ become scalars, and the regularization penalty, $\sum_{g=1}^{G}\gamma_g^\top\Lambda\gamma_g$, becomes equivalent to $\lambda\sum_{g=1}^{G}\gamma_g^2$ from Equation (9). Equation (11), however, allows varying slopes. For example, if $Z_{g[i]} = [1\ X^{(1)}_{g[i]}]^\top$, then $\gamma_g = [\gamma_{0g}\ \gamma_{1g}]^\top$ and the regularization penalty becomes $\sum_{g=1}^{G}\gamma_g^\top\Lambda\gamma_g = \sum_{g=1}^{G}(\lambda_{00}\gamma_{0g}^2 + 2\lambda_{01}\gamma_{0g}\gamma_{1g} + \lambda_{11}\gamma_{1g}^2)$ where λ_{00} and λ_{11} make up the diagonal Λ and λ_{01} is its off-diagonal element.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

MLM can produce more accurate (out-of-sample) outcome predictions than does FE: regularization prevents the over-fitting that FE may allow, particularly in the case of small groups.[15]

Further, MLM is sometimes understood to "wisely" adapt the level of shrinkage based on group size, but the comparison to regularization shows that such adaptation is not as sophisticated as it may appear. When MLM or regFE encounters a group that appears to have a very high or low mean relative to others, choosing an extreme γ_g would decrease the sum of squared errors for that group, $\sum_{i=1}^{n_g}(Y_{g[i]} - X_{g[i]}^\top \beta - Z_{g[i]}^\top \gamma_g)^2$. For a relatively small group, the savings in squared loss would be smaller relative to the additional "cost" paid through the regularization penalty, $\gamma_g^\top \Lambda \gamma_g$, and so γ_g will be left relatively close to zero. By comparison, when MLM or regFE encounters a larger group, the savings in terms of squared loss would justify the added regularization penalty, so the choice of γ_g minimizing the penalized sum of squared errors will be a more extreme one. Although this behavior appears to intelligently balance prior knowledge with allowing the data to speak, it is reproduced simply through regularization.

Another feature of MLM that can be understood through the regularization view is its ability to estimate coefficients for group-level variables even while including group-specific intercepts and slopes that would have prevented model identification under OLS. This occurs for the same reason that one can include more coefficients than observations in a "ridge regression." Consider attempting to find $(\hat{\beta}, \hat{\gamma})$ purely by OLS,

$$(\hat{\beta}, \hat{\gamma}) = \underset{\beta,\gamma}{\arg\min} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} [Y_{g[i]} - X_{g[i]}^\top \beta - Z_{g[i]}^\top \gamma_g]^2 \right) = \underset{\beta,\gamma}{\arg\min} \left\| Y - \begin{bmatrix} X & Z \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \right\|_2^2. \tag{12}$$

This has no unique solution if $X_{g[i]}$ contains a group-level variable or its interaction with a variable in $Z_{g[i]}$, as $[X\ Z]^\top [X\ Z]$ is then singular. However, by introducing the regularization penalty into the minimization problem, regFE essentially adds to $[X\ Z]^\top [X\ Z]$ a positive semi-definite matrix that allows the sum of the two matrices to be invertible.

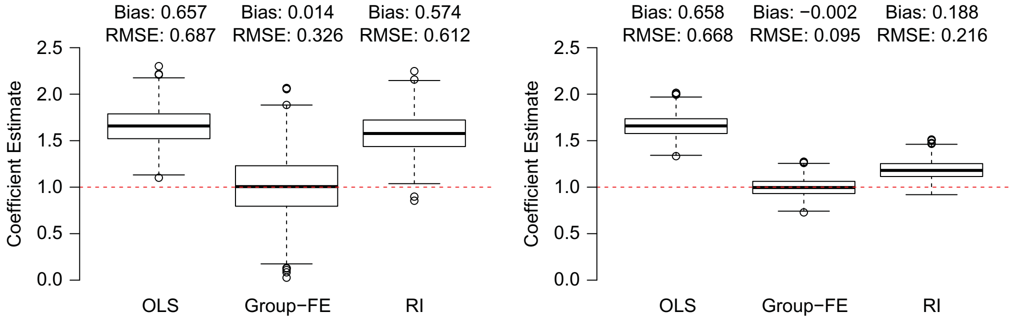
Finally, this equivalence illuminates the main concern with MLM: bias when the random effects are correlated with $X_{g[i]}$. Because the group-specific intercepts in a RI model are regularized, they do not achieve the values that would "fully absorb" group-specific confounding, leaving components unexplained that can instead be captured by biasing the coefficients on $X_{g[i]}$. In Equation (2), where we hope to condition on group to absorb group-level confounders, random effects thus offer only "incomplete conditioning," not fully accounting for the unobserved group-level variables' influence on the outcome. To illustrate, consider the following DGP:

$$\begin{aligned} & Y_{g[i]} = \beta_0 + \beta_1 X_{g[i]} + (W_g^{(1)} + W_g^{(2)}) + \epsilon_{g[i]} \\ \text{where} \quad & [W_g^{(1)}\ W_g^{(2)}]^\top \overset{iid}{\sim} \mathcal{N}(0, 2\mathcal{I}_2)\,,\ X_{g[i]} = W_g^{(1)} + N(0,1)_{g[i]}\,,\ \epsilon_{g[i]} \overset{iid}{\sim} N(0, \sigma^2), \end{aligned} \tag{DGP1}$$

where $X_{g[i]} \in \mathbb{R}$ is an observed observation-level variable and the $W_g^{(\ell)}$ are unobserved group-level covariates, with $W_g^{(1)}$ being a confounder. While Group-FE will unbiasedly estimate β_1, a simple OLS regression of Y on X would produce a biased estimate of β_1, having failed to account for $W_g^{(1)}$. RI may seem more appealing than OLS because investigators might hope that the γ_g will capture the group-level confounding. However, the shrinkage of γ_g due to treating them as random effects yields estimated intercepts closer to zero than those obtained by Group-FE and required to

[15] One well-known use of MLM, especially in political science, is "multilevel regression and post-stratification" (MRP; Park, Gelman, and Bafumi 2004). This is an approach to small-area estimation, in which estimates for the conditional means of small groups are attempted despite having very little data by group. Its strength in this task is derived from the ability to partially pool information across units, that is, shrinkage of the random effect estimates. Other approaches that explicitly engage regularization may also thus be effective in this task.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press



(a) $G = 50 \quad n_g = 5$ (b) $G = 50 \quad n_g = 50$

Figure 2. Comparison of estimates of β_1 from OLS, Group-FE, and RI in DGP1. *Note:* Results across 1000 iterations, each drawn from DGP 1 with $\beta_0 = \beta_1 = 1$. The dashed-line represents the true β_ℓ. Due to correlated random effects, RI estimates are almost as biased as OLS estimates when group size is small (5). The bias is less severe but still appreciable at a group size of 50, and RMSE remains twice that of Group-FE.

account for $W_g^{(1)}$. This leaves some of $W_g^{(1)}$ unabsorbed, allowing it to continue biasing $\hat{\beta}_1$ as in the OLS model. We call this “incomplete conditioning” because the intended analytical strategy required to estimate β unbiasedly would condition on group, but the use of RI fails to achieve this.

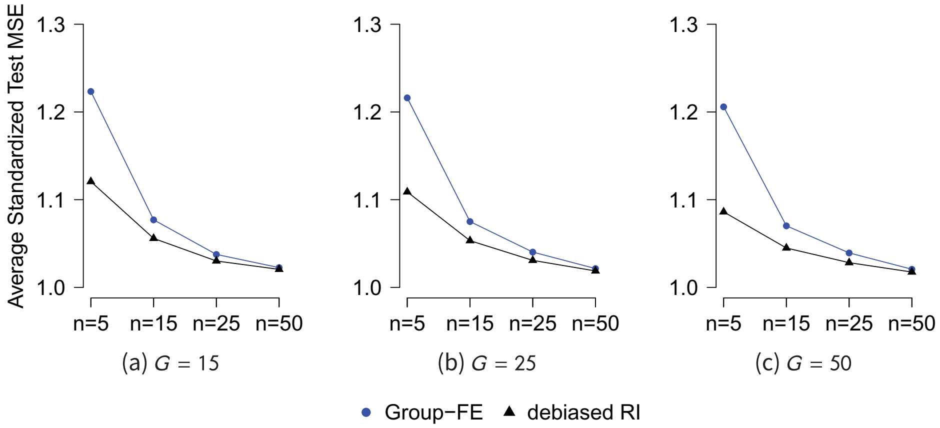
Figure 2 illustrates this. Define bias and root mean square error (RMSE) as,

$$\text{Bias} = \frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)} - \beta), \qquad \text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)} - \beta)^2}$$

where m indexes the number of iterations from 1 to M and $\hat{\beta}^{(m)}$ is the estimate from the mth iteration. Bias is large, as expected, for OLS as it fails to account for group-level confounding at all. RI makes almost no improvement when groups are small ($n_g = 5$), and only a partial improvement when the groups are quite large ($n_g = 50$). Note that unbiasedness of RI would require the absence of correlated random effects, that is no correlation between $X_{g[i]}$ and $W_g^{(1)}$, which is to say an absence of group-level confounding. Had this been the case, OLS would also suffice, and would differ from RI only in its efficiency and how variance is estimated. By contrast, in the presence of such correlation, Group-FE effectively eliminates confounding bias at both group sizes. In terms of efficiency, while RI has the expected slight decrease in variance, its RMSE remains about twice that of Group-FE at either group size due to these biases. That RI's average bias falls as n_g rises may seem to be a cause for hope when one has a large enough dataset, and Lockwood, McCaffrey, *et al.* (2007) describes the conditions under which this type of bias tends to 0 as $n_g \to \infty$ generally. However, in practice, there is no knowing if n_g is large enough to ensure negligible bias in a given case, as this depends on the correlation between the covariates and the random effects. We also note that with group-level covariates, increasing n_g may not alleviate bias (see Appendix A.4).

Comparison to practice. For each of the three analyses in Sections 3.1, 3.2 and 3.3, we briefly contrast what is already known of these claims to what we find in practice. In this case, both textbooks and pedagogical articles remark heavily on the correlated random effects assumption, albeit not usually in terms of regularization or incomplete conditioning. Yet, this most central of concerns regarding MLM is demonstrably neglected in empirical practice. Among the MLM-based studies we reviewed where such bias would be at issue, only one in 24 education articles, 13 of 39 political science articles, and 10 of 19 sociology articles addressed the issue.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press



(a) $G = 15$ (b) $G = 25$ (c) $G = 50$

Figure 3. Outcome prediction error for debiased RI versus Group-FE. *Note:* Comparison of testing error for the predicted outcome (average standardized test MSE, $(N\mathbb{E}(\epsilon^2_{g[i]}))^{-1}\sum_{g,i}(Y_{g[i]}-\hat{Y}_{g[i]})^2)$. The RI model with $\Sigma = \sigma^2 I_N$ has been debiased by including $\bar{X}_g$ as a covariate, and shows lower testing error, especially when groups are smaller. Results are averaged across 1000 iterations, each drawn from DGP 1. Testing data are of the same size as the training data.

3.2 Bias-corrected MLM

The second analytical piece is that while MLM's bias problem with correlated random effects appears to be dire, there is a simple solution dating back to Mundlak (1978). With this fix, MLM unbiasedly estimates coefficients for variables below the level of grouping or clustering, like FE does, while retaining the ability to include group-level variables in the model and the superior predictive accuracy that arises from regularization/random effects. For RI, the fix requires adding to the model the group-level means of $X_{g[i]}$ (including any interactions or other nonlinear terms).

As a stepping stone, consider a similar approach that could be applied to OLS as a substitute for FE. We consider again DGP 1 from Section 3.1, but suppose we estimate the following model by OLS,

$$Y_{g[i]} = \beta_0 + \beta_1 X_{g[i]} + \alpha_1 \bar{X}_g + \epsilon_{g[i]}, \tag{13}$$

where $\bar{X}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} X_{g[i]}$. Informally, adding $\bar{X}_g$ "soaks up" any contribution that $W_g^{(1)}$ could have made to $Y_{g[i]}$ that could be correlated with $X_{g[i]}$, protecting the estimate of β_1 in the same way that FE would, and leaving an unbiased estimate of β_1 under the conditional independence assumption. Proof is given in Appendix A.5. This bias-curing effect of including $\bar{X}_g$ in OLS can similarly be applied to the RI model, and alleviates RI's bias problem, with $\Sigma = \sigma^2 I_N$, in estimating β_1. We omit the proof of this result, as we prove a more general claim below. In fact, an OLS model including $\bar{X}_g$, the Group-FE model, and the RI model including $\bar{X}_g$ all produce exactly the same point estimates. Despite this equivalence, the RI model retains MLM's greater (out-of-sample) predictive accuracy for the outcome compared to Group-FE—a benefit of the regularization imposed on γ_g, as illustrated in Figure 3.

One can easily generalize this alteration to other models with varying intercepts but with multiple covariates or treatments: simply add to the RI model the group-level means of all included variables, $\bar{X}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} X_{g[i]}$, including any interactions or nonlinear transformations. If spherical observation-level errors are assumed, the coefficient estimates from this model for lower-level variables are exactly equal to those obtained by Group-FE. Historically, the inclusion of $\bar{X}_g$ was proposed by Mundlak (1978) and extended by Chamberlain (1979, 1982), and is known in some econometrics-informed traditions as the "correlated random effects" (CRE) approach (Wooldridge

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

2010; Schunck 2013). Although originally motivated by imposing the assumption $\mathbb{E}(\gamma_g \mid X_g, Z_g) = \bar{X}_g^\top \alpha$, we show unbiasedness without this assumption by showing its equivalence to "partialing out" of the group-level effects, which is in turn equivalent to Group-FE. A closely related approach is the hybrid model of Allison (2009), which group-demeans $X_{g[i]}$ in addition to including $\bar{X}_g$. This model is an isomorphic variation on models that only include $\bar{X}_g$ without demeaning, producing the same coefficient of interest on X but altering the way in which coefficients are combined to interpret between-group differences (see Schunck 2013).

We now turn to the more general debiasing approach for cases allowing multiple random coefficients and not just group intercepts. We refer to this as "bias-corrected MLM" (**bcMLM**), and use this label throughout the remainder of the paper to cover the special case of bias-corrected RI as well.[16] We first take the projections of the "fixed effect variables" (X_g), excluding the intercept, onto the random effect variables (Z_g) within each group, $\tilde{X}_{g[i]} = Z_{g[i]}^\top (Z_g^\top Z_g)^{-1} Z_g^\top X_g$.[17] These projections are then added to the regression as "fixed effect variables,"

$$Y_{g[i]} = X_{g[i]}^\top \beta + \tilde{X}_{g[i]} \alpha + Z_{g[i]}^\top \gamma_g + \epsilon_{g[i]}, \quad \gamma_g \mid X, Z \overset{iid}{\sim} \mathcal{N}(0, \Omega) \qquad \text{(bcMLM)}$$

where β and α are assumed fixed, and we continue to assume that $\epsilon_g \mid X, Z \overset{ind}{\sim} \mathcal{N}(0, \Sigma_g)$ and $\epsilon_{g[i]} \perp\!\!\!\perp \gamma_{g'} \mid X, Z$ for all g, g', and i. When $\Sigma = \sigma^2 I_N$, and provided it is not overidentified (described below), bcMLM produces estimates for β that are unbiased under the conditional independence assumption (Appendix A.8) and identical to FE estimates (Appendix A.9).

We make two remarks on this result. First, it provides unbiasedness only when spherical observation-level errors are assumed.[18] While the choice of variance-covariance matrix for the errors has no impact on point estimates under FE, the point estimates from MLM are sensitive to this choice. Second, the RI model including $\bar{X}_g$ noted above is a special case of bcMLM: if $Z_g = \vec{\mathbb{1}}_{n_g}$, then $\tilde{X}_{g[i]} = [1](\vec{\mathbb{1}}_{n_g}^\top \vec{\mathbb{1}}_{n_g})^{-1} \vec{\mathbb{1}}_{n_g}^\top X_g = \bar{X}_g^\top$.

Limitations and the per-cluster regression. One limitation of bcMLM is that it can fail to eliminate potential biases for coefficients of certain variables because including $\tilde{X}_{g[i]}$ results in an overidentified model. A simple example would be a RI model that includes a group-level covariate, U_g, in $X_{g[i]}$. The proposed alteration suggests that one includes $\bar{U}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} U_g$, but this is simply the same U_g already included (see Appendix A.6 for an example). Therefore, unless U_g is independent of the random effects and any included lower-level variables (e.g., if U_g were randomly assigned as a group-level treatment), the estimated coefficients for U_g may be biased.[19]

Thankfully, as long as U_g is uncorrelated with the random intercepts, its coefficient can be unbiasedly estimated if desired by adding a "per-cluster regression" step as proposed by Bates *et al.* (2014). For example, suppose we have one observation-level variable $X_{g[i]}$ and one group-level variable U_g, with coefficients β_1 and β_2, respectively. First, using Group-FE or bcMLM (here, an RI model including $\bar{X}_g$), unbiasedly estimate $\hat{\beta}_1$. Then remove the estimated marginal effect of X from Y, forming $Y_{g[i]}^\perp = Y_{g[i]} - \hat{\beta}_1 X_{g[i]}$. The per-cluster step is to then regress the G group-level means of $Y_{g[i]}^\perp$ on U_g and an intercept term by OLS. We provide an example of this process in Appendix A.12.

[16] Similar suggestions have been made by Wooldridge (2005), Snijders and Berkhof (2008), and Raudenbush (2009). Another approach (Snijders and Bosker 2011; Wooldridge 2013), is to include in an MLM the interaction of $(\bar{X}_g - \bar{X})$ with all variables in $Z_{g[i]}$, where $\bar{X} = \sum_{i,j} X_{g[i]}$ is the grand-mean of $X_{g[i]}$. This amounts to including $(\bar{X}_g - \bar{X})^\top \otimes Z_{g[i]}$ among the $X_{g[i]}$, where $\otimes$ is the Kronecker product. However, this is not guaranteed to debias estimates of β (see Appendix A.10 and A.11).

[17] Note that this is a slight abuse of notation, as X_g contains a column for the intercept term. In practice, this should be removed from X_g when forming these projections. We appear to keep it here in the interest of avoiding extra notation.

[18] We are not aware of the general unbiasedness or consistency of bcMLM when $\Sigma \neq \sigma^2 I_N$ and is possibly misspecified.

[19] Furthermore, if $\text{cor}(X_{g[i]}^{(\ell)}, U_g) \neq 0$ for some lower-level variable $X_{g[i]}^{(\ell)}$, the inclusion of $\bar{X}_g^{(\ell)}$ may induce bias in the coefficient for U_g that would otherwise not be there (demonstrated in Appendix A.7).

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

Another important example of an over-identification limitation to bcMLM occurs when the slope for $X^{(\ell)}_{g[i]}$ is allowed to vary, that is, $X^{(\ell)}_{g[i]}$ is included in $Z_{g[i]}$. Because bcMLM includes as extra covariates the predictions of $X^{(\ell)}_{g[i]}$ using $Z_{g[i]}$, and $Z_{g[i]}$ predicts $X^{(\ell)}_{g[i]}$ perfectly, including this "prediction" simply includes $X^{(\ell)}_{g[i]}$ in the model twice. One of the $X^{(\ell)}_{g[i]}$ would be dropped out of the model, and the "prediction" of $X^{(\ell)}_{g[i]}$ cannot soak up the bias from any potentially correlated random effects when estimating β_ℓ. This is also true of coefficients for any cross-level interactions, $X^{(\ell)}_{g[i]}U_g$. Here again the per-cluster regression provides an option for users who are interested in those coefficients, as demonstrated in Appendix A.13.[20]

Comparison to practice. The main takeaway from this analysis is that bcMLM removes the bias of MLM due to correlation of random effects with the treatment, and in so doing, produces coefficient estimates identical to FE (when $\Sigma = \sigma^2 I_N$ is assumed), while retaining MLM's superior predictive accuracy for the outcome and the ability to model group-level variables, which may or may not be of interest to investigators depending on their task. Yet, among articles we reviewed where bias due to correlated random effects would be at issue, none of 24 education articles, one in 39 political science articles, and 2 of 19 sociology articles employed bcMLM or any other suitable debiasing approach for MLM.

3.3 Variance estimation

The third and final analytical piece is that the "default" MLM standard errors are based on narrow assumptions that are often inappropriate, but this too can be remedied. A commonly cited motive for using MLM is the claim that it ensures appropriate standard errors for multilevel data (e.g., Luke 2004; Morgan and Kelly 2017; Beazer and Blake 2018; Campbell and Ronfeldt 2018; Rueda 2018; Clements *et al.* 2019; Pardos-Prado and Xena 2019). This is only true when the sole source of heteroskedasticity or dependency between residuals arises due to the random effects included by $Z_{g[i]}$ and their contributions to $Y_{g[i]}$. That a default standard error exists for MLM, and is the only choice in some software, does not imply it is always an appropriate choice.

Recall from Section 2.4 that the random effects contribution, $Z^\top_{g[i]}\gamma_g$, and the idiosyncratic error, $\epsilon_{g[i]}$, can be thought of as a single mean-zero error term $\epsilon^*_{g[i]} = Z^\top_{g[i]}\gamma_g + \epsilon_{g[i]}$ in the model $Y_{g[i]} = X^\top_{g[i]}\beta + \epsilon^*_{g[i]}$. The dependency between two observations' outcomes within the same group, conditional on X and Z, is then

$$\text{cov}(Y_{g[i]}, Y_{g[i']} \mid X, Z) = \text{cov}(\epsilon^*_{g[i]}, \epsilon^*_{g[i']} \mid X, Z) = Z^\top_{g[i]}\Omega Z_{g[i']} + \text{cov}(\epsilon_{g[i]}, \epsilon_{g[i']} \mid X, Z). \quad (14)$$

MLM can be understood as a framework for structuring this covariance, by specifying which variables enter the models as random effects (i.e., $Z_{g[i]}$) and parameterizing Ω and Σ. In the RI model, with $Z_{g[i]} = [1]$, the combined error, $\epsilon^*_{g[i]}$, is $\gamma_g + \epsilon_{g[i]}$. Under the conditional independence assumption and spherical errors, $\epsilon_{g[i]} \mid X, Z \overset{iid}{\sim} N(0, \sigma^2)$, the dependence structure is

$$\text{var}(Y_{g[i]} \mid X, Z) = \omega^2 + \sigma^2 \quad \text{and} \quad \text{cov}(Y_{g[i]}, Y_{g[i']} \mid X, Z) = \omega^2 \quad \text{for} \quad i \neq i' \quad (15)$$

In other words, $Y_{g[i]}$ is modeled as linear in $X_{g[i]}$ with error, but instead of independent observations, there is constant covariance between observations in the same group, and this covariance does not differ by group. This yields a compound symmetric covariance matrix for each group's

[20] The per-cluster regression does, however, require that all $n_g > d$, and is unstable when any non-intercept elements of $Z_{g[i]}$ have little variation within a group. Graham and Powell (2012), extending a closely related estimator from Chamberlain (1992), had previously investigated the conditions under which β is identifiable in these cases despite correlated random effect contributions, and proposes an estimator that is consistent when they hold.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

error vector, $\epsilon_g^* = Y_g - X_g\beta$. If $Z_{g[i]} = [1 \; X_{g[i]}^{(1)}]^\top$, then $\epsilon_{g[i]}^* = \gamma_{0g} + \gamma_{1g} X_{g[i]}^{(1)} + \epsilon_{g[i]}$. Maintaining that $\epsilon_{g[i]} \mid X, Z \overset{iid}{\sim} N(0, \sigma^2)$ and that (conditionally on X and Z) γ_{0g} and γ_{1g} are drawn from $N(0, \omega_0^2)$ and $N(0, \omega_1^2)$ with covariance ω_{01} yields intragroup covariances of

$$\begin{aligned} \mathrm{var}(Y_{g[i]} \mid X, Z) &= \omega_0^2 + 2X_{g[i]}^{(1)}\omega_{01} + \left[X_{g[i]}^{(1)}\right]^2 \omega_1^2 + \sigma^2 \\ \text{and} \quad \mathrm{cov}(Y_{g[i]}, Y_{g[i']} \mid X, Z) &= \omega_0^2 + \left[X_{g[i]}^{(1)} + X_{g[i']}^{(1)}\right]\omega_{01} + X_{g[i]}^{(1)} X_{g[i']}^{(1)} \omega_1^2 \quad \text{for} \quad i \neq i' \end{aligned} \tag{16}$$

With the addition of more random effect variables, the variance structure becomes more complex. This complexity should not be equated with generality—the variance is still assumed to be a highly prescribed function of the data. To illustrate the potential for variance estimates with poor coverage, consider the following longitudinal DGP, where g indexes the person and $t = 1, \ldots, T$ indexes the time-point of the observation:

$$\begin{aligned} & Y_{g[t]} = \beta_0 + \beta_1 X_{g[t]} + \beta_2 U_g + W_g + \epsilon_{g[t]} \\ \text{where} \quad & W_g \overset{iid}{\sim} N(0, 4), \\ & X_{g[t]} \sim N(0, 1) \quad \text{and} \quad \mathrm{cor}(X_{g[t]}, X_{g[t+k]}) = (0.75)^k, \\ & U_g \overset{iid}{\sim} N(0, 1), \; \epsilon_{g[i]} \sim N(0, U_g^2\sigma^2) \quad \text{and} \quad \mathrm{cor}(\epsilon_{g[t]}, \epsilon_{g[t+k]}) = (0.75)^k \end{aligned} \tag{DGP2}$$

Here, there is an observation-level variable $X_{g[t]}$ that is auto-correlated; a group-level variable U_g; and a random intercept W_g. The observation-level error terms are auto-correlated as well with (heteroskedastic) variances that depend on U_g. The correct dependence structure would be

$$\mathrm{cov}(Y_{g[t]}, Y_{g[t+k]} \mid X, Z) = U_g^2\sigma^2(0.75)^k + 4 \tag{17}$$

Using the "default" variance with RI, assuming $\Sigma = \sigma^2 I_N$, Figure 4 shows coverage rates for nominal 95% confidence intervals of β_1 and β_2 across draws from DGP 2. Typically we would focus interest on β_1, motivated by an assumption of no within-group confounding. However, we also show results for β_2 because group-level variables were often of interest in the empirical works we reviewed. The RI standard errors are consistently too small for both β_1 and β_2 across all sample sizes. The undercoverage for β_1 worsens as the total number of time-periods (T) increases. Coverage improves for β_2 as T increases, but remains unsatisfactory even at $T = 50$. Similar undercoverage of the "default" MLM standard errors for RI has been noted by Bell, Fairbrother, and Jones (2019), Heisig, Schaeffer, and Giesecke (2017), and Jacqmin-Gadda *et al.* (2007).

Relaxing assumptions for MLM variance estimation. If the user has strong reason to believe errors (conditionally on the random effect contributions) are spherical, then the default MLM standard errors just described would be appropriate. However, such justifications are rarely offered. Fortunately, more flexible approaches can be employed in the MLM framework, relaxing this assumption. For example, with longitudinal data, it is possible to assume an AR(1) error structure for the $\epsilon_{g[t]}$, in which $\mathrm{cor}(\epsilon_{g[t]}, \epsilon_{g[t+k]}) = \rho^k$ for $\rho \in (-1, 1)$. Or, one may allow $\mathrm{var}(\epsilon_{g[i]} \mid X, Z) = \sigma_g^2$, where σ_g^2 can differ by group, to accommodate heteroskedasticity by group. One may also allow a common unstructured Σ_g across g, which makes no assumptions on the intragroup covariance and assigns a separate parameter to each $\mathrm{cov}(\epsilon_{g[i]}, \epsilon_{g[i']})$.[21]

[21] Another avenue toward achieving flexibility is to allow the random effects to be heteroskedastic, that is, relaxing the assumption that the γ_g have common variance, Ω. Hoffman (2015), for example, proposes directly modeling the variance of the random effects as a function of the covariates, such as $\mathrm{var}(\gamma_{g0} \mid X, Z) = \exp(\nu_0 + \nu_1 U_g)$ where U_g is a group-level covariate and (ν_0, ν_1) are parameters to be estimated.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

Using such specialized error structures when estimating standard errors may be advisable when researchers can defend the corresponding assumptions. On the other hand, users often cannot claim to know the correct variance structure based on theory alone. Fortunately, cluster-robust standard errors (CRSE), popular in conjunction with FE, provide a useful low-assumption alternative by asking the user only to assume independence across groups while allowing within group covariance to be fully estimated, albeit at the cost of requiring more data.[22]

Note that both MLM and CRSEs operate as though we assume zero covariance of the residuals from units in different groups, sharing the assumption

$$\mathrm{cov}(Y_{g[i]}, Y_{g'[i']} \mid X, Z) = \begin{cases} \mathbb{E}(\epsilon^*_{g[i]}\epsilon^*_{g'[i']} \mid X, Z) & \text{if } g = g' \\ 0 & \text{if } g \neq g' \end{cases} \tag{18}$$

What differs is only how $\mathbb{E}(\epsilon^*_{g[i]}\epsilon^*_{g[i']} \mid X, Z)$ is determined. In MLM, this covariance is parametrized as described above (e.g., $\mathbb{E}[\epsilon^*_{g[i]}\epsilon^*_{g[i']} \mid X, Z] = \omega^2 + \mathbb{1}_{\{i=i'\}}\sigma^2$ in a RI model), and estimated by plugging in the parameter estimates. By contrast CRSEs are remarkable for the lack of structure they impose on these within-group covariances. For example, after estimating $\hat{\beta}$ with an OLS of Y on X, CRSEs would simply construct empirical covariance estimates

$$\hat{\mathbb{E}}_{\mathrm{CRSE}}(\epsilon^*_{g[i]}\epsilon^*_{g'[i']} \mid X, Z) = \begin{cases} c \times \hat{e}_{g[i]}\hat{e}_{g'[i']} & \text{if } g = g' \\ 0 & \text{if } g \neq g' \end{cases} \tag{19}$$

where $\hat{e}_{g[i]} = Y_{g[i]} - X^{\top}_{g[i]}\hat{\beta}$ and c is a scalar finite sample correction. In other words, while MLMs make a strict assumption on the within-group covariances, CRSEs impose among the weakest possible assumptions by simply employing an empirical estimate based on fitted residuals. Appendix A.14 provides a more detailed discussion of the CRSE structure.

Fortunately, nothing prevents MLM users from employing CRSE assumptions during variance estimation (see also Cameron and Miller 2015), estimating variance according to

$$\widehat{\mathrm{var}}_{\mathrm{CRSE}}(\hat{\beta}_{\mathrm{MLM}}) = c \times (X^{\top}\hat{V}^{-1}_{\mathrm{MLM}}X)^{-1}X^{\top}\hat{V}^{-1}_{\mathrm{MLM}}\begin{bmatrix} \hat{e}_1\hat{e}_1^{\top} & & 0 \\ & \ddots & \\ 0 & & \hat{e}_G\hat{e}_G^{\top} \end{bmatrix}\hat{V}^{-1}_{\mathrm{MLM}}X(X^{\top}\hat{V}^{-1}_{\mathrm{MLM}}X)^{-1} \tag{20}$$

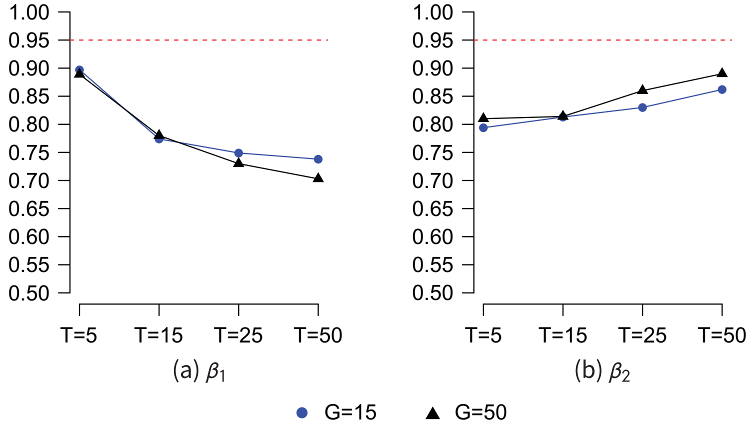
where $\hat{e}_g = Y_g - X_g\hat{\beta}_{\mathrm{MLM}}$ and V is defined in Equation (4). We discuss the choice of c in Appendix A.15.

This leads to a suprising and useful equivalance. Mirroring the equivalence between estimates of β from bcMLM with $\Sigma = \sigma^2 I_N$ and FE, *the CRSEs forβ from both models are also equal* if both use the same c (as we recommend in Appendix A.15). This fact, proven in Appendix A.16, may be surprising since the point estimates of $\hat{Y}_{g[i]}$ differ between models, and thus the overall error variance differs. This equivalence avoids debate over which method is appropriate to estimate the coefficient and standard error on the covariate of interest, since the answers will be the same.

Figure 5 illustrates the performance of CRSE with MLM in DGP 2, in which RI with $\Sigma = \sigma^2 I_N$ misspecifies the dependence structure. Confidence intervals for β_1 using CRSEs fixes the undercoverage seen above in Figure 4 using the conventional standard errors. Coverage remains poor for β_2 with $G = 15$, with some undercoverage remaining at $G = 50$.

[22] We refer readers to Cameron and Miller (2015) for a detailed review of CRSEs. We take a model-based perspective here and show the dependence structure implied by different variance estimators, including CRSEs. For a design-based approach to considering when clustering may be required and concerns regarding the conservative affects of clustering at too high a level, see Abadie *et al.* (2017).

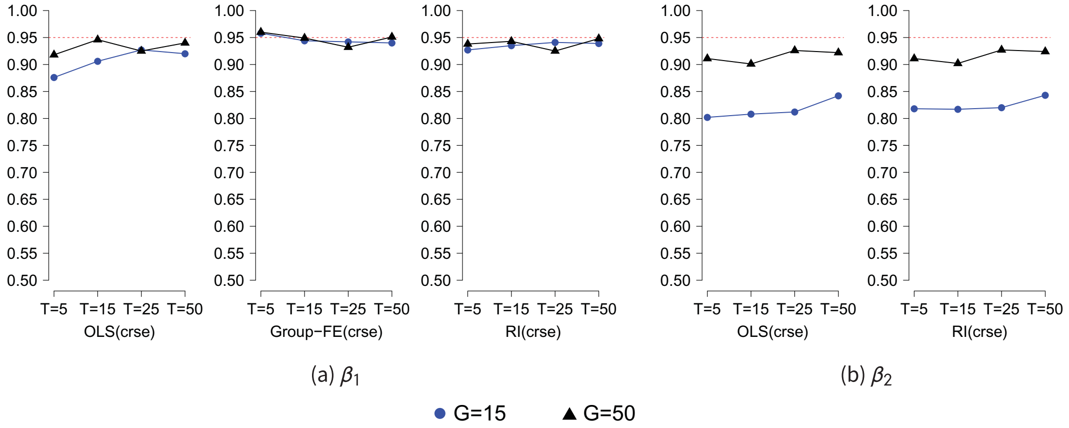
https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press



(a) β_1

(b) β_2

Figure 4. Coverage rates under RI, assuming $\Sigma = \sigma^2 I_N$ in DGP 2. *Note:* Coverage rates for 95% nominal confidence intervals (vertical axis) for β_1 (*left*) and β_2 (*right*). Results across 1000 iterations, each drawn from DGP 2 with $\beta_0 = \beta_1 = \beta_2 = 1$. The dashed-line represents the target coverage rate of 0.95.



(a) β_1

(b) β_2

Figure 5. Coverage rates of 95% confidence intervals from RI, OLS, and Group-FE, all with CRSE, in DGP 2. *Note:* Results across 1000 iterations, each drawn from DGP 2, with $\beta_0 = \beta_1 = \beta_2 = 1$.

Guidance on CRSEs. We offer several notes regarding the appropriate use of CRSEs. First, CRSEs assume that observations in different groups have zero covariance in their residuals. Investigators must keep this in mind when choosing the level at which clustering is performed.[23] Clustering units that actually belong to different groups as if they are in the same group reduces the number of clusters but does not violate the CRSE assumption. By contrast, if units that actually have dependent residuals are labeled as if they belong to separate groups, the CRSE assumption of no between-group dependence will be violated and the results will be unreliable. In some cases, there may not be any choice of grouping that makes this assumption defensible, in which case CRSEs would not be defensible either.

Second, as in any modeling problem, there is a tradeoff between the ability to relax assumptions and the requirement for more data. Thus while CRSEs can be a substantial improvement over default RI model-based standard errors, they do so at the cost of demanding more data. The convergence of the CRSEs depends on the number of groups. Cameron and Miller (2015) suggests that 20 to 50 clusters may be needed to ensure stable estimates. Naturally, this number depends upon many features of the data and no guidelines can be expected to be universally sufficient. Alternatively, when investigators can arguably defend the stricter assumptions of any less flexible

[23] This logic extends naturally to data with multi-way clustering (e.g., clustering by both time and country in country-year data) by assuming arbitrary dependence between any two observations that are grouped together on any dimension, and no dependence between observations that are not grouped on any dimension (see e.g., Cameron and Miller 2015).

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

covariance structure, such as an AR(1) or a simpler heteroskedastic model for example, then doing so may pay off. To this end, when the number of groups is smaller, the model-based MLM standard errors may be preferable if one can justify that the assumed dependence structure is plausible.

Comparison to practice. Most textbooks we reviewed describe alternative estimators for the variance of MLM such as auto-regressive models.[24] Empirical works using MLM we reviewed showed little attention to this issue: among articles that employed RI models to estimate a coefficient of interest, all 24 articles in education, 29 of 39 articles in political science articles, and 14 of 21 articles in sociology used the default MLM standard errors ($\Sigma = \sigma^2 I_N$) without discussion or justification. We surmise there are several reasons for this. The first is the misconception, documented above, that MLM automatically produces correct standard errors for any multilevel data structure without further consideration. Second and compounding the first, software widely used for MLM estimation does not always allow alternative variance structures besides that with $\Sigma = \sigma^2 I$ or nonconstant Ω.[25] Finally, investigators may reasonably worry that correctly specifying covariance structures using theory or prior knowledge is not feasible, and decide to instead accept default choices. This makes the CRSE approach a particularly attractive option, at least when groups can be defined such that between-group residual dependence is arguably ruled out.

4 Conclusions

Different methodological traditions have responded to the challenges posed by grouped data with divergent solutions: FE with modified standard errors, or MLMs with random effects. To demystify their properties, we show that (i) random effects invoked in MLMs can be understood as regularized FE, explaining MLM's improved predictive power, ability to include group-level variables, and bias problem; (ii) this bias can be addressed; and (iii) the "default" standard errors under MLM do not necessarily address all concerns with intragroup dependency in multilevel data.

We thus recommend estimating coefficients with either FE or bcMLM, with the assumption of spherical errors. In both cases, CRSEs offer a flexible approach to variance estimation, particularly if an argument can be made for independence across clusters. Fortunately, these two approaches produce identical point estimates and standard errors for the coefficients they share. Hence, for those willing to make these adjustments and focused on inference regarding a treatment coefficient, the question of whether FE or MLM is "more appropriate" is irrelevant. Both approaches sacrifice the potential efficiency gain that an uncorrected MLM would offer *had its strict assumptions been true*.[26] We consider this a small and acceptable price to pay to avoid the risk of bias and higher RMSE that occurs under uncorrected MLM in the presence of correlated random effects. Accordingly, we suggest these unbiased approaches rather than any approach that seeks to mix estimators (e.g., Cheng, Liao, and Shi 2019) or to choose between FE (or bcMLM) and uncorrected MLM based on some criterion. For example, we do not advocate for choosing uncorrected MLM when the number of observations per group is above some threshold: one cannot know how many observations will be enough for the bias (and RMSE) to become acceptably small, and any potential efficiency or accuracy gain of MLM relative to FE is diminishing in group size anyway. We similarly do not advocate for a statistical testing approach such as Hausman (1978): if one is concerned that

[24] Only Snijders and Bosker (2011) discussed CRSEs in depth. Among pedagogical articles, Cameron and Miller (2015) clearly describe the connection between MLM standard errors and CRSE, while Heisig, Schaeffer, and Giesecke (2017) compare coverage rates of model-based MLM standard errors and CRSEs in simulations.

[25] At the time of writing, `lme4` in `R` and `VARCOMP` in `SPSS` do not allow nonspherical observation-level errors, while `nlme` in `R`, `SAS MIXED`, and `MIXED` in `SPSS` do. `SAS NLMIXED` also allows random effects to be heteroskedastic. Alternatively, more general Bayesian modeling and sampling software such as WinBUGS and STAN allow very flexible models.

[26] bcMLM may have efficiency gains over FE, however, if one specifies a different model for Σ that nearly enough approximates the correct structure. See Appendix A.2. However, we are not aware of a proof of the general unbiasedness or consistency of bcMLM when nonspherical errors are assumed but Σ is possibly misspecified.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

one of those estimates (from uncorrected random effects) is incorrect, then knowing whether the observed difference in the estimates is statistically significant or not is of little relevance.

Although our advice regarding CRSEs is less common, our debiasing recommendations echo Raudenbush (2009), Bell and Jones (2015), and Bell, Fairbrother, and Jones (2019). It is also consistent with the "correlated random effects" approaches in econometrics, which employ the Mundlak (1978) solution, noted in texts including Wooldridge (2010) and Greene (2012). Nevertheless, such advice have gone largely unheeded in political science and education, and to some degree in sociology.[27]

Finally, while FE and bcMLM produce identical results for the coefficients they share, the approaches differ in that (i) bcMLM has improved predictive (out-of-sample) accuracy for the outcome, and (ii) bcMLM retains the ability to include group-level covariates and cross-level interactions in the model. Whether users are interested in predictive accuracy from the same model in which they are interested in estimating an unbiased effect of a key covariate is a question of research goals, not addressed here. We also emphasize that the estimated coefficients for group-level covariates or cross-level interactions in bcMLM may be difficult to interpret. In addition to the usual identification concerns, the bias-correction step in bcMLM applies only to coefficients it shares with FE, and may even *induce* bias in those that are absent from FE. For users interested in these coefficients, bcMLM or FE regression can be followed by the appropriate per-cluster regression step.

Acknowledgements

We thank Michael Bates, Neal Beck, Jennie Brand, Kevin Esterling, Jeff Gill, Jeff Lewis, Jingyi Jessica Li, Ian Lundberg, Michael Seltzer, Brandon Stewart, attendees of the 2019 Southern California Political Methodology meeting, and anonymous reviewers for their valuable comments and suggestions.

Data Availability Statement

Replication materials are available at Hazlett and Wainstein (2020).

Supplementary Material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2020.41.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge. 2017. When Should You Adjust Standard Errors for Clustering ? *Technical Report*. National Bureau of Economic Research.

Allison, P. D. 2009. *Fixed Effects Regression Models*. Vol. 160. Los Angeles, CA: SAGE Publications.

Bates, M. D., K. E. Castellano, S. Rabe-Hesketh, and A. Skrondal. 2014. "Handling Correlations Between Covariates and Random Slopes in Multilevel Models." *Journal of Educational and Behavioral Statistics* 39(6):524–549.

Beazer, Q. H., and D. J. Blake. 2018. "The Conditional Nature of Political Risk: How Home Institutions Influence the Location of Foreign Direct Investment." *American Journal of Political Science* 62(2):470–485.

Bell, A., M. Fairbrother, and K. Jones. 2019. "Fixed and Random Effects Models: Making an Informed Choice." *Quality & Quantity* 53(2):1051–1074.

Bell, A., and K. Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3(1):133–153.

Bollen, K. A., and J. E. Brand. 2010. "A General Panel Model with Random and Fixed Effects: A Structural Equations Approach." *Social Forces* 89(1):1–34.

[27] Other fields have better internalized this advice, particularly psychology, as also noted by McNeish and Kelley (2018) and apparent in several methodological texts including Kim and Frees (2007) and Hox and Roberts (2011). We also note that Bollen and Brand (2010) introduced another approach to handling correlated random effects that involves estimating the correlation between the random effects and $X_{g[i]}$ within a structural equation modeling framework.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

Cameron, A. C., and D. L. Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50(2):317–372.

Campbell, S. L., and M. Ronfeldt. 2018. “Observational Evaluation of Teachers: Measuring More Than We Bargained for?” *American Educational Research Journal* 55(6):1233–1267.

Chamberlain, G. 1979. *Analysis of Covariance with Qualitative Data*. Cambridge, MA: National Bureau of Economic Research.

Chamberlain, G. 1982. “Multivariate Regression Models for Panel Data.” *Journal of Econometrics* 18(1):5–46.

Chamberlain, G. 1992. “Efficiency Bounds for Semiparametric Regression.” *Econometrica: Journal of the Econometric Society* 60:567–596.

Cheng, X., Z. Liao, and R. Shi. 2019. “On Uniform Asymptotic Risk of Averaging GMM Estimators.” *Quantitative Economics* 10(3):931–979.

Clark, T. S., and D. A. Linzer. 2015. “Should I Use Fixed or Random Effects?” *Political Science Research and Methods* 3(2):399–408.

Clements, D. H., J. Sarama, A. J. Baroody, C. Joswick, and C. B. Wolfe. 2019. “Evaluating the Efficacy of a Learning Trajectory for Early Shape Composition.” *American Educational Research Journal* 56(6):2509–2530.

Czado, C. 2017. “*Lecture 10: Linear Mixed Models (Linear Models with Random Effects)*.”

Faraway, J. J. 2016. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton, FL: Chapman & Hall/CRC Press.

Finch, W. H., J. E. Bolin, and K. Kelley. 2016. *Multilevel Modeling Using R*. Boca Raton, FL: CRC Press.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware. 2004. *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Gelman, A. 2006. “Multilevel (Hierarchical) Modeling: What It Can and Cannot Do.” *Technometrics* 48(3): 432–435.

Gelman, A. 2005. “Analysis of Variance—Why It Is More Important Than Ever.” *The Annals of Statistics* 33(1):1–53.

Gelman, A., and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Graham, B. S., and J. L. Powell 2012. “Identification and Estimation of Average Partial Effects in “Irregular” Correlated Random Coefficient Panel Data Models.” *Econometrica* 80(5):2105–2152.

Greene, W. H. 2003. *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. 2012. *Econometric Analysis*. New York: Prentice Hall.

Hausman, J. A. 1978. “Specification Tests in Econometrics.” *Econometrica: Journal of the Econometric Society* 46:1251–1271.

Hazlett, C., and L. Wainstein. “Replication data for: Understanding, choosing, and unifying multilevel and fixed effect approaches.” https://doi.org/10.7910/DVN/VZDPSQ, Harvard Dataverse, V1.

Heck, R. H., S. L. Thomas, and L. N. Tabata. 2013. *Multilevel and Longitudinal Modeling with IBM SPSS*. New York: Routledge.

Heisig, J. P., M. Schaeffer, and J. Giesecke. 2017. “The Costs of Simplicity: Why Multilevel Models May Benefit from Accounting for Cross-Cluster Differences in the Effects of Controls.” *American Sociological Review* 82(4):796–827.

Hoffman, L. 2015. *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change*. London: Routledge.

Hox, J., and J. K. Roberts. 2011. *Handbook of Advanced Multilevel Analysis*. New York: Psychology Press.

Imai, K., and I. S. Kim. 2019. “When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?” *American Journal of Political Science* 63(2):467–490.

Jacqmin-Gadda, H., S. Sibillot, C. Proust, J.-M. Molina, and R. Thiébaut. 2007. “Robustness of the Linear Mixed Model to Misspecified Error Distribution.” *Computational Statistics & Data Analysis* 51(10):5142–5154.

Kim, J.-S., and E. W. Frees. 2007. “Multilevel Modeling with Correlated Effects.” *Psychometrika* 72(4):505–533.

Kim, Y., and P. Steiner. “Causal Graphical Views of Fixed Effects and Random Effects Models.” 2019.

Lockwood, J., D. F. McCaffrey. 2007. “Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement.” *Electronic Journal of Statistics* 1:223–252.

Luke, D. A. 2004. *Multilevel Modeling*, Vol. 143. Los Angeles, CA: SAGE Publications.

McElreath, R. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: Chapman & Hall/CRC Press.

McNeish, D., and K. Kelley. 2018. “Fixed Effects Models Versus Mixed Effects Models for Clustered Data: Reviewing the Approaches, Disentangling the Differences, and Making Recommendations.” *Psychological Methods* 24:20.

Morgan, J., and N. J. Kelly. 2017. “Social Patterns of Inequality, Partisan Competition, and Latin American Support for Redistribution.” *The Journal of Politics* 79(1):193–209.

Mundlak, Y. 1978. “On the Pooling of Time Series and Cross Section Data.” *Econometrica: Journal of the Econometric Society* 46(1):69–85.

Pardos-Prado, S., and C. Xena. 2019. “Skill Specificity and Attitudes Toward Immigration.” *American Journal of Political Science* 63(2):286–304.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press

Park, D. K., A. Gelman, and J. Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–385.

Pearl, J. 2000. *Causality: Models, Reasoning and Inference*, Vol. 29. New York: Springer.

Raudenbush, S. W. 2009. "Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings." *Education Finance and Policy* 4(4):468–491.

Rueda, D. 2018. "Food Comes First, Then Morals: Redistribution Preferences, Parochial Altruism, and Immigration in Western Europe." *The Journal of Politics* 80(1):225–239.

Schunck, R. 2013. "Within and Between Estimates in Random-Effects Models: Advantages and Drawbacks of Correlated Random Effects and Hybrid Models." *The Stata Journal* 13(1):65–76.

Snijders, T. A., and J. Berkhof. 2008. "Diagnostic Checks for Multilevel Models." In *Handbook of Multilevel Analysis*, edited by J. de Leeuw and E. Meijer, 141–175. New York: Springer.

Snijders, T. A., and R. J. Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Los Angeles, CA: SAGE Publications.

Steenbergen, M. R., and B. S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46(1):218–237.

White, H. 1984. "Asymptotic Theory for Econometricians." *Technical report*.

White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–838.

Wooldridge, J. M. 2005. "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models." *Review of Economics and Statistics* 87(2):385–390.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. 2013. *Correlated Random Effects Panel Data Models*. IZA Summer School in Labor Economics. http://conference.iza.org/conference_files/SUMS_2013/slides_1_linear_iza.pdf.

https://doi.org/10.1017/pan.2020.41 Published online by Cambridge University Press