Online Supplement

Understanding, choosing, and unifying multilevel and fixed effect approaches

Chad Hazlett & Leonard Wainstein

A Appendix

A.1 Tables with symbols and abbreviations

Table 1. Symbols

Symbol	Description	Relevant model(s)	Location(s)
α	Coefficient vector	bcMLM	Section 3.2
β	Coefficient vector	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
c	Scalar	FE, Group-FE, MLM, RI, bcMLM	Section 3.3 Appendix A.15
$\hat{e}_{g[i]}$	Random (residual) variable	FE, MLM, bcMLM	Section 3.3
$\hat{e}_g$ and $\hat{e}$	Random (residual) vector	FE, MLM, bcMLM	Section 3.3 Appendix A.16
$\epsilon_{g[i]}$	Random (error) variable	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
ϵ_g and ϵ	Random (error) vector	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
$\epsilon^*_{g[i]}$	Random (error) variable	MLM, RI, bcMLM	Section 3.3
ϵ^*_g and ϵ^*	Random (error) vector	MLM, RI, bcMLM	Section 2.4
γ_g and γ	Coefficient vector	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
ω^2	Scalar variance	RI	Section 2.2
Ω and Ω_{block}	Covariance matrix	MLM, bcMLM	Section 2.2
λ	Scalar tuning parameter	regFE	Section 3.1
Λ	Matrix tuning parameter	regFE	Section 3.1
σ^2	Scalar variance	MLM, RI, bcMLM	Section 2.2
Σ_g and Σ	Covariance matrix	MLM, RI, bcMLM	Section 2.2
V_g and V	Covariance matrix	MLM, RI, bcMLM	Section 2.4 Appendix A.8 Appendix A.16
$X_{g[i]}$	Random (covariate) vector	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
X_g and X	Random (covariate) matrix	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
$\bar{X}_g$	Random (covariate) vector	bcMLM	Section 3.2
$\tilde{X}_{g[i]}$	Random (covariate) vector	FE, bcMLM	Section 3.2 Appendix A.8 Appendix A.16
$\tilde{X}_g$ and $\tilde{X}$	Random (covariate) matrix	FE, bcMLM	Section 3.2 Appendix A.8 Appendix A.16
$Y_{g[i]}$	Random (outcome) variable	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
Y_g and Y	Random (outcome) vector	FE, Group-FE, MLM, RI, regFE, bcMLM	Section 2.1
$Z_{g[i]}$	Random (covariate) vector	FE, MLM, regFE, bcMLM	Section 2.1
Z_g and Z	Random (covariate) matrix	FE, MLM, regFE, bcMLM	Section 2.1

Table 2. Abbreviations for model-related terms

Abbreviation	Full name	Location(s)
bcMLM	Bias-corrected multilevel model	Section 3.2
CRSE	Cluster-robust standard error	Section 3.3 Appendix A.14
FE	Fixed effects model	Section 2.2
Group-FE	Group fixed effects model	Section 2.2
MLM	Multilevel model	Section 2.2
regFE	Regularized fixed effects model	Section 3.1
RI	Random intercepts model	Section 2.2

A.2 Extensive simulation

We present here an example in which applying the lessons of Section 3 allows users to navigate a complicated data generating process. We consider a longitudinal setting in which

$$Y_{g[t]} = \beta_0 + \beta_1 X^{(1)}_{g[t]} + \beta_2 U^{(1)}_g + \beta_3 X^{(1)}_{g[t]} U^{(1)}_g + (5W^{(1)}_g + 5W^{(1)}_g W^{(2)}_g) + \epsilon_{g[t]} \qquad \text{(DGP 3)}$$

where $[W^{(1)}_g \; W^{(2)}_g]^\top \overset{iid}{\sim} N(0, 2I_2)$,

$$X^{(1)}_{g[t]} = W^{(1)}_g + W^{(2)}_g + \delta_{g[t]} \quad \text{where} \quad \delta_{g[t]} \sim N(0,1) \quad \text{and} \quad \text{cor}(\delta_{g[t]}, \delta_{g[t+k]}) = (0.75)^k,$$

$$U^{(1)}_g = W^{(2)}_g + N(0,1)_g,$$

$$\epsilon_{g[i]} \sim N(0, [U^{(1)}_g]^2 \sigma^2) \quad \text{and} \quad \text{cor}(\epsilon_{g[t]}, \epsilon_{g[t+k]}) = (0.75)^k$$

where $W^{(1)}_g$ and $W^{(2)}_g$ are unobserved and $t = 1, \ldots, T$, with T varying from 5 to 50 in different settings. Like DGP 1, the random intercept, $(5W^{(1)}_g + 5W^{(1)}_g W^{(2)}_g)$, is correlated with the covariates, specifically the observation-level variable, $X^{(1)}_{g[t]}$, and the cross-level interaction, $X^{(1)}_{g[t]} U^{(1)}_g$, threatening to bias at least β_1 and β_3. Like DGP 2, the dependence structure is complex:

$$\text{cov}(Y_{g[t]}, Y_{g[t+k]} \mid X, Z) = \text{var}(5W^{(1)}_g + 5W^{(1)}_g W^{(2)}_g \mid X, Z) + [U^{(1)}_g]^2 \sigma^2 (0.75)^k \qquad (22)$$

which shows that covariance arises not only due to the random intercept but also autocorrelation in $\epsilon_{g[t]}$, with varying intensity by group.

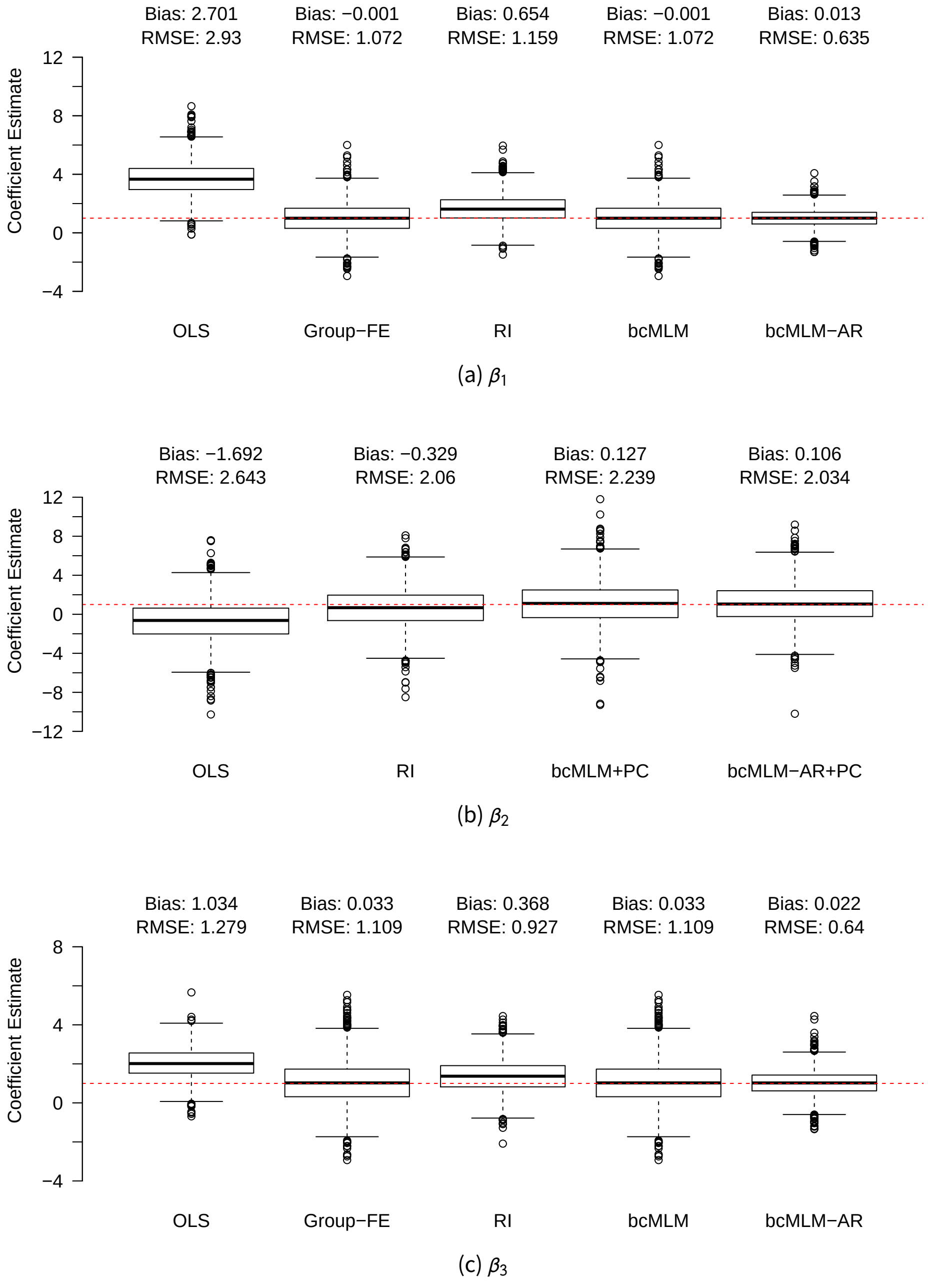
One choice would be to employ bcMLM through a RI model that additionally includes $\bar{X}^{(1)}_g$ and $\bar{X}^{(1)}_g U^{(1)}_g$. We consider both choices of $\Sigma = \sigma^2 I_N$ (bcMLM) and an AR(1) structure for each Σ_g with constant variances (bcMLM-AR), to attempt to capture the longitudinal nature of the data.

Additionally, though we emphasize that coefficients on group-level variables or cross-level interactions are often not clearly linked to causal quantities of direct interest, they are commonly spoken of in practice. To this end we also employ the per-cluster regression (*model*+PC, when relevant) to estimate β_2 after bcMLM and bcMLM-AR (note that $5W^{(1)}_g + 5W^{(1)}_g W^{(2)}_g$ is uncorrelated with $U^{(1)}_g$, while $U^{(1)}_g$ and $X^{(1)}_{g[t]}$ are correlated, so the per-cluster regression should estimate β_2 unbiasedly and bcMLM may not). We will compare these to a simple OLS of Y on X (OLS), Group-FE, and a RI model without bias-correction and $\Sigma = \sigma^2 I_N$ (RI). For standard errors, we use model-based MLM standard errors (Section 2.4; *model*(mlm), when relevant) with RI, use CRSEs (Section 3.3; *model*(crse), when relevant) with bcMLM, OLS, and Group-FE, and try both variance estimators with bcMLM-AR.[28] Furthermore, when employing the per-cluster regression, we use robust standard errors (White *et al.* 1980; *model*(robust)) from the group-level regression in the final step of the procedure.

The bias, coverage rates, and average standardized test mean square error for each of these models is shown in Figures 6, 7, and 8, each across choices of $T \in \{5, 15, 25, 50\}$ and $G \in \{15, 50\}$. Due to space limitations, we only show bias plots for the scenario in which $G = 50$ and $T = 25$. However, the pattern of results is similar across all sample sizes tried.

[28] Note that bcMLM-AR still misspecifies the dependence structure, as the idiosyncratic errors are heteroskedastic in truth, so we should expect its model-based standard errors to be incorrect.

Figure 6. Comparison of five models on DGP 3: Estimating β_1, β_2, and β_3



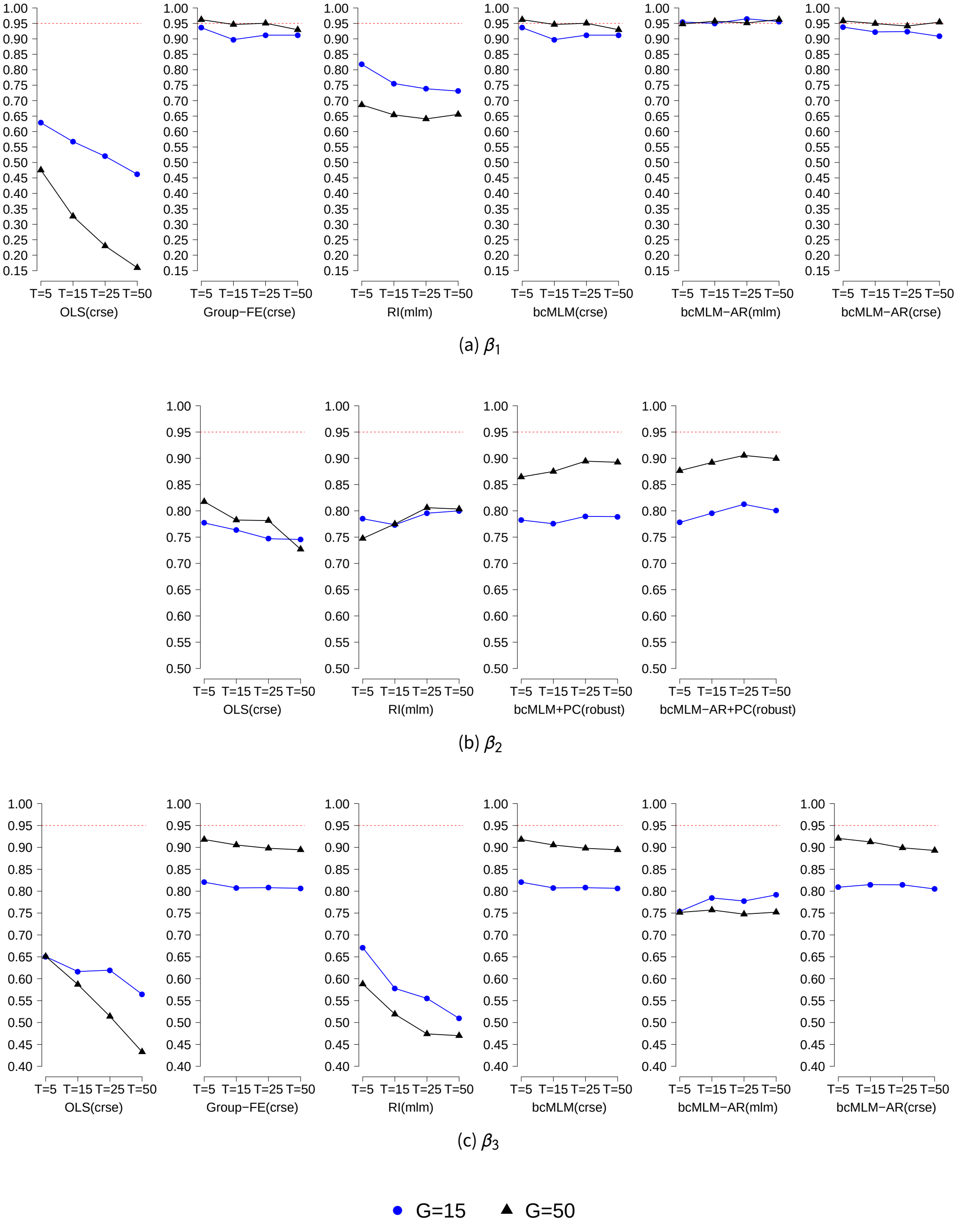
(a) β_1

(b) β_2

(c) β_3

Note: Results across 2000 iterations, each drawn from DGP 3 with $G = 50$, $T = 25$, and $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$. The red dashed-line represents the true β_ℓ.

Figure 7. Comparison of five models on DGP 3: Coverage of β_1, β_2, and β_3



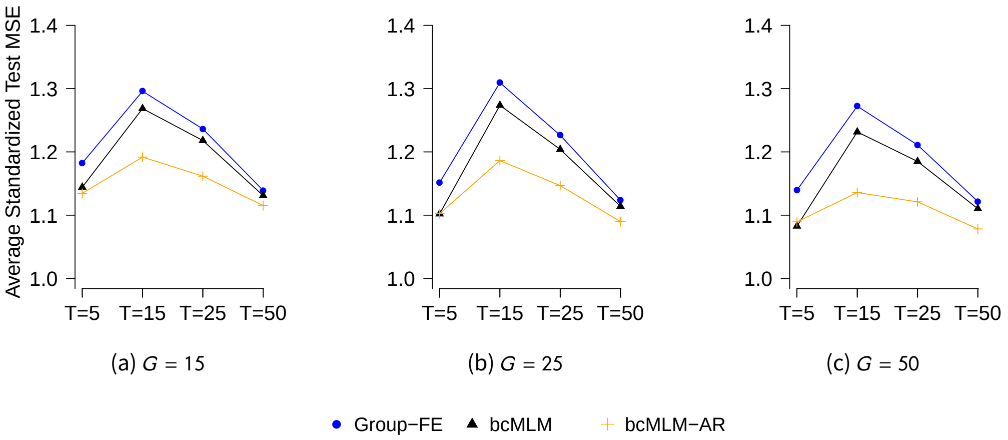
(a) β_1

(b) β_2

(c) β_3

Note: Results across 2000 iterations, each drawn from DGP 3 with $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$.

Figure 8. Outcome prediction error for Group-FE, bcMLM, and bcMLM-AR in DGP 3



(a) $G = 15$

(b) $G = 25$

(c) $G = 50$

Note: Comparison of testing error for the predicted outcome (average standardized test MSE, $(N\mathbb{E}(\epsilon^2_{g[i]}))^{-1}\sum_{g,i}(Y_{g[i]} - \hat{Y}_{g[i]})^2$). Results are averaged across 2000 iterations, each drawn from DGP 3 with $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$. Testing data are of the same size as are the sample data. Additionally, due to the longitudinal nature of DGP 3, testing data are for time points immediately after those of the sample data (e.g., when $T = 5$, the sample data span $1 \le t \le 5$ and testing data span $6 \le t \le 10$).

Regarding bias and RMSE, bcMLM-AR(+PC), bcMLM(+PC), and Group-FE all show no bias for β_1, β_2, and β_3. Though RI has slightly lower RMSE for β_2 and β_3 than do bcMLM(+PC) and Group-FE, this comes at the cost of noticeable bias for most coefficients. OLS also shows severe bias. bcMLM-AR(+PC) and bcMLM(+PC) perform equally well in terms of bias for each coefficient, but bcMLM-AR(+PC) produces noticeably more efficient estimates (lower RMSE) than do bcMLM(+PC) or Group-FE, likely because the AR(1) structure for Σ_g more nearly resembles the true structure than does $\Sigma_g = \sigma^2 I_{n_g}$.

Turning to coverage, bcMLM(crse), bcMLM-AR(crse), and Group-FE(crse) all show imperfect but perhaps acceptable coverage rates for β_1, and consistently show undercoverage for β_3, particularly when $G = 15$. bcMLM+PC(robust) and bcMLM-AR+PC(robust) also consistently show undercoverage for β_2. bcMLM-AR(mlm) shows acceptable coverage for β_1, but is outperformed by bcMLM(crse), bcMLM-AR(crse), and Group-FE(crse) for β_3. OLS(crse) and RI(mlm) perform poorly for all coefficients due the models' biased estimates, and in RI(mlm)'s case, a grossly misspecified dependence structure.

As for predictive accuracy, bcMLM-AR and bcMLM are uniformly superior to Group-FE. bcMLM-AR largely performs better than bcMLM, especially when T is larger, likely due to the former's more efficient coefficient estimates. That the average standardized test mean square error for the three models first increases from $T = 5$ to $T = 15$ before steadily decreasing as T increases is likely due to the autocorrelation in both $\epsilon_{g[t]}$ and $X^{(1)}_{g[t]}$, and should not be expected in other DGPs.

Overall, these models perform as expected: bcMLM is equivalent to Group-FE, and together with bcMLM-AR these models are clearly the best for estimating β_1 and β_3. The per-cluster approach is effective in recovering β_2 from a bias point of view, but provides poor coverage. OLS and RI show substantial biases, failing to account for the group level confounding. Additionally, though more DGP-dependent, bcMLM-AR has an efficiency advantage over bcMLM while showing equally low bias. The bias-corrected MLMs also have superior predictive accuracy over Group-FE. Finally, bcMLM(crse), bcMLM-AR(crse), and bcMLM-AR(mlm) all achieve acceptable coverage for β_1. But given the poor coverage of bcMLM-AR(mlm) on β_3, bcMLM(crse) and bcMLM-AR(crse) have the best overall performance, at least when G is larger.

A.3 Proof of Theorem 3.2

We prove this by applying properties of $\hat{\beta}_{\text{MLM}}$ and $\hat{\gamma}_{\text{MLM}}$ that are proven in Czado 2017. We then reframe these properties into the context of regFE to show the equivalence. As shown in Section 2.4, given $(\hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}})$, and subsequently $\hat{V}_{\text{MLM}}$ by substituting for Ω and Σ,

$$\hat{\beta}_{\text{MLM}} = (X^\top \hat{V}_{\text{MLM}}^{-1} X)^{-1} X^\top \hat{V}_{\text{MLM}}^{-1} Y \tag{23}$$

$$\hat{\gamma}_{\text{MLM}} = \begin{bmatrix} \hat{\Omega}_{\text{MLM}} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega}_{\text{MLM}} \end{bmatrix} Z^\top \hat{V}_{\text{MLM}}^{-1} (Y - X\hat{\beta}_{\text{MLM}}) \tag{24}$$

We remind readers that $\hat{\beta}_{\text{MLM}}$ is found before $\hat{\gamma}_{\text{MLM}}$, specifically by maximizing the likelihood of β given $Y, Z, X, \hat{\Omega}_{\text{MLM}}$, and $\hat{\Sigma}_{\text{MLM}}$,

$$\begin{aligned} \hat{\beta}_{\text{MLM}} &= \arg\max_{\beta} L(\beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}} \mid Y, X, Z) \\ &= \arg\max_{\beta} p(Y \mid X, Z, \beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}}) \end{aligned} \tag{25}$$

And $\hat{\gamma}_{\text{MLM}}$ is subsequently found by maximizing the posterior distribution of γ given $Y, Z, X, \hat{\Omega}_{\text{MLM}}$, $\hat{\Sigma}_{\text{MLM}}$ and $\hat{\beta}_{\text{MLM}}$,

$$\hat{\gamma}_{\text{MLM}} = \arg\max_{\gamma} p(\gamma \mid Y, X, Z, \hat{\beta}_{\text{MLM}}, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}}) \tag{26}$$

Now, consider instead an alternate procedure that estimates β and γ simultaneously by maximizing the joint distribution of Y and γ:

$$\begin{aligned} &\arg\max_{\beta,\gamma} p(Y, \gamma \mid X, Z, \beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}}) \\ =&\arg\max_{\beta,\gamma} p(Y \mid \gamma, X, Z, \beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}}) p(\gamma \mid X, Z, \beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}}) \\ =&\arg\max_{\beta,\gamma} \Big(\log p(Y \mid \gamma, X, Z, \beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}}) + \log p(\gamma \mid X, Z, \beta, \hat{\Omega}_{\text{MLM}}, \hat{\Sigma}_{\text{MLM}})\Big) \end{aligned} \tag{27}$$

Because $Y \mid \gamma, X, Z \sim \mathcal{N}(X\beta + Z\gamma, \Sigma)$ and $\gamma \mid X, Z \sim \mathcal{N}(0, \begin{bmatrix} \Omega & & 0 \\ & \ddots & \\ 0 & & \Omega \end{bmatrix})$, the problem becomes

$$\begin{aligned} &\arg\max_{\beta,\gamma} \Big(-\frac{1}{2}(Y - X\beta - Z\gamma)^\top \hat{\Sigma}_{\text{MLM}}^{-1} (Y - X\beta - Z\gamma) - \frac{1}{2}\gamma^\top \begin{bmatrix} \hat{\Omega}_{\text{MLM}}^{-1} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega}_{\text{MLM}}^{-1} \end{bmatrix} \gamma \Big) \\ =&\arg\min_{\beta,\gamma} \Big((Y - X\beta - Z\gamma)^\top \hat{\Sigma}_{\text{MLM}}^{-1} (Y - X\beta - Z\gamma) + \gamma^\top \begin{bmatrix} \hat{\Omega}_{\text{MLM}}^{-1} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega}_{\text{MLM}}^{-1} \end{bmatrix} \gamma \Big) \end{aligned} \tag{28}$$

Let Q be the objective function in the above minimization problem, i.e.,

$$\begin{aligned}
&Q(\beta, \gamma \mid \hat{\Sigma}_{\mathrm{MLM}}, \hat{\Omega}_{\mathrm{MLM}}) \\
&\quad = (Y - X\beta - Z\gamma)^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} (Y - X\beta - Z\gamma) + \gamma^{\top} \begin{bmatrix} \hat{\Omega}^{-1}_{\mathrm{MLM}} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega}^{-1}_{\mathrm{MLM}} \end{bmatrix} \gamma \qquad (29)
\end{aligned}$$

Minimizing Q for β and γ involves finding β and γ that satisfy $\frac{\partial Q}{\partial \beta} = 0$ and $\frac{\partial Q}{\partial \gamma} = 0$. It is easily found that this amounts to

$$\frac{\partial Q}{\partial \beta} = X^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} X\beta + X^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} Z\gamma - X^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} Y = 0 \qquad (30)$$

$$\frac{\partial Q}{\partial \gamma} = Z^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} X\beta + \left(Z^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} Z + \begin{bmatrix} \hat{\Omega}^{-1}_{\mathrm{MLM}} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega}^{-1}_{\mathrm{MLM}} \end{bmatrix} \right) \gamma - Z^{\top} \hat{\Sigma}^{-1}_{\mathrm{MLM}} Y = 0 \qquad (31)$$

One finds that substituting $\beta = \hat{\beta}_{\mathrm{MLM}}$ and $\gamma = \hat{\gamma}_{\mathrm{MLM}}$ satisfies the above equations. Therefore,

$$(\hat{\beta}_{\mathrm{MLM}}, \hat{\gamma}_{\mathrm{MLM}}) = \underset{\beta, \gamma}{\arg\min}\, Q(\beta, \gamma \mid \hat{\Sigma}_{\mathrm{MLM}}, \hat{\Omega}_{\mathrm{MLM}}) \qquad (32)$$

The results presented so far can be found in Czado 2017. Further inspection of Q leads to the equivalence between MLM and regFE in the theorem. Letting $\Sigma = \sigma^2 I_N$ as in the theorem, maximizing Q is equivalent to

$$\begin{aligned}
&\underset{\beta, \gamma}{\arg\min}\, Q(\beta, \gamma \mid \hat{\Sigma}_{\mathrm{MLM}}, \hat{\Omega}_{\mathrm{MLM}}) \\
&= \underset{\beta, \gamma}{\arg\min} \left(\hat{\sigma}^{-2}_{\mathrm{MLM}} ||Y - X\beta - Z\gamma||_2^2 + \gamma^{\top} \begin{bmatrix} \hat{\Omega}^{-1}_{\mathrm{MLM}} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega}^{-1}_{\mathrm{MLM}} \end{bmatrix} \gamma \right) \\
&= \underset{\beta, \gamma}{\arg\min} \left(||Y - X\beta - Z\gamma||_2^2 + \gamma^{\top} \begin{bmatrix} \hat{\sigma}^2_{\mathrm{MLM}} \hat{\Omega}^{-1}_{\mathrm{MLM}} & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}^2_{\mathrm{MLM}} \hat{\Omega}^{-1}_{\mathrm{MLM}} \end{bmatrix} \gamma \right) \\
&= \underset{\beta, \gamma}{\arg\min} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} [Y_{g[i]} - X^{\top}_{g[i]} \beta - Z^{\top}_{g[i]} \gamma_g]^2 + \sum_{g=1}^{G} \gamma_g^{\top} (\hat{\sigma}^2_{\mathrm{MLM}} \hat{\Omega}^{-1}_{\mathrm{MLM}}) \gamma_g \right) \qquad (33)
\end{aligned}$$

Letting $\Lambda = \hat{\sigma}^2_{\mathrm{MLM}} \hat{\Omega}^{-1}_{\mathrm{MLM}}$ as in the theorem, making this substitution leads to the exact minimization problem for regFE,

$$\underset{\beta, \gamma}{\arg\min} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} [Y_{g[i]} - X^{\top}_{g[i]} \beta - Z^{\top}_{g[i]} \gamma_g]^2 + \sum_{g=1}^{G} \gamma_g^{\top} \Lambda \gamma_g \right) \qquad (34)$$

So, given the conditions for the theorem, $\hat{\beta}_{\mathrm{MLM}}$ and $\hat{\gamma}_{\mathrm{MLM}}$ solve the minimization problem for regFE, giving the equivalence.

□

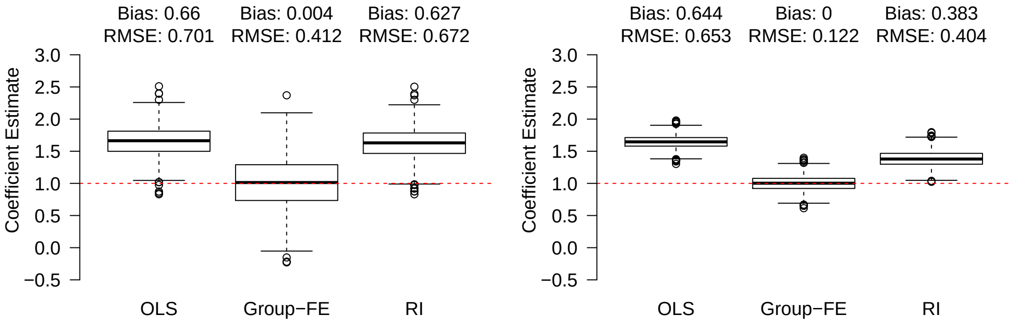
A.4 Simulated example: biased $\hat{\beta}_{\text{MLM}}$ for group-level variables in the presence of correlated random effects

Consider the following DGP:

$$\begin{aligned} Y_{g[i]} &= \beta_0 + \beta_1 X^{(1)}_{g[i]} + \beta_2 U^{(1)}_g + (W^{(1)}_g + W^{(2)}_g) + \epsilon_{g[i]} \qquad \text{(DGP 4)} \\ \text{where} \quad [W^{(1)}_g \; W^{(2)}_g]^\top &\overset{iid}{\sim} \mathcal{N}(0, 2\mathcal{I}_2), \\ X^{(1)}_{g[i]} &= W^{(1)}_g + N(0,1)_{g[i]} \\ U^{(1)}_g &= W^{(2)}_g + N(0,1)_g \\ \epsilon_{g[i]} &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

Here, there is an observed lower-level variable, $X^{(1)}_{g[i]}$, and an observed group-level variable, $U^{(1)}_g$, which are both correlated with the unobserved random intercept $(W^{(1)}_g + W^{(2)}_g)$. Comparing the analogous OLS, Group-FE, and RI models in draws from this DGP with $\beta_0 = \beta_1 = \beta_2 = 1$, we again see, in Figures 9 and 10, consistent biases in estimates of β_1 and β_2 from the RI model.

Figure 9. Comparison of estimates of β_1 from OLS, Group-FE, and RI in DGP 4



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

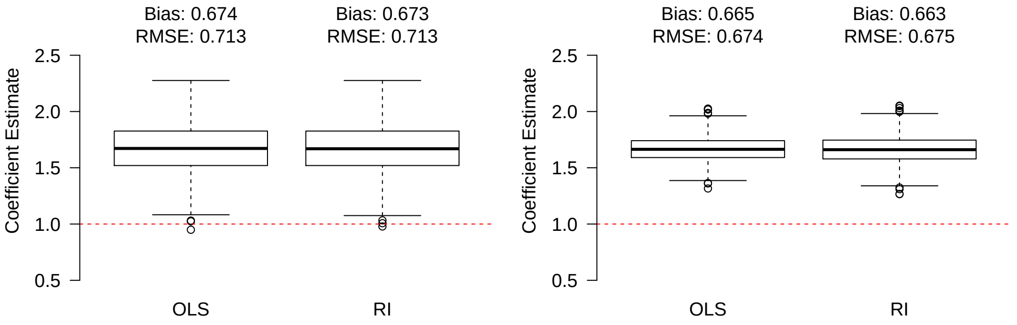
Note: Results across 1000 iterations, each drawn from DGP 4 with $\beta_0 = \beta_1 = \beta_2 = 1$. The red dashed-line represents the true β_ℓ.

RI is between OLS and Group-FE in terms of bias in estimating β_1, and improves as n_g increases just like it does in DGP 1. However, in terms of estimating β_2, RI is just as poor as is OLS at both choices of n_g, and does not improve as n_g increases. This is because the conditional mean of $W^{(2)}_g$ is linear in $U^{(1)}_g$, so that portion of $(W^{(1)}_g + W^{(2)}_g)$ is explained just as well by $U^{(1)}_g$ as the included random intercept, γ_g, regardless of n_g or G. So, MLM automatically chooses $U^{(1)}_g$ over γ_g to explain that portion of $Y_{g[i]}$ due to the shrinkage imposed on γ_g.

A.5 Proof of the unbiasedness of an OLS including $\bar{X}_g$ in DGP 1

This result stems from an equivalence between the $\hat{\beta}_1$ from an OLS including $\bar{X}_g$ (Equation (13)) and that from a regression of $Y_{g[i]}$ on $(X_{g[i]} - \bar{X}_g)$, which is also unbiased (and, in fact, the same as that from a Group-FE model). Letting $X^{\perp}_{g[i]} = X_{g[i]} - \bar{X}_g$ and $W_g = W^{(1)}_g + W^{(2)}_g$, a $\hat{\beta}_1$ from a

Figure 10. Comparison of estimates of β_2 from OLS and RI in DGP 4



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: Results across 1000 iterations, each drawn from DGP 4 with $\beta_0 = \beta_1 = \beta_2 = 1$. The red dashed-line represents the true β_ℓ.

regression of $Y_{g[i]}$ on $X^{\perp}_{g[i]}$ yields (by the FWL theorem)

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{g,i} Y_{g[i]}(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} \\
&= \frac{\sum_{g,i}\Big(\beta_0 + \beta_1 X_{g[i]} + W_g + \epsilon_{g[i]}\Big)(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} \\
&= \frac{\sum_{g,i}\Big(\beta_0 + \beta_1 (X_{g[i]} - \bar{X}_g + \bar{X}_g) + W_g + \epsilon_{g[i]}\Big)(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} \\
&= \frac{\sum_{g,i}\Big(\beta_1 X^{\perp}_{g[i]} + (\beta_0 + \beta_1 \bar{X}_g + W_g) + \epsilon_{g[i]}\Big)(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} \\
&= \beta_1 + \frac{\sum_{g,i}(\beta_0 + \beta_1 \bar{X}_g + W_g)(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} + \frac{\sum_{g,i}\epsilon_{g[i]}(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} \qquad (35)
\end{aligned}$$

$\sum_{g,i}(\beta_0 + \beta_1 \bar{X}_g + W_g)(X^{\perp}_{g[i]}) = 0$ because within each group, $\bar{X}_g$ and W_g are constant and $X^{\perp}_{g[i]}$ is mean-zero. More rigorously,

$$\begin{aligned}
\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\beta_0 + \beta_1 \bar{X}_g + W_g)(X^{\perp}_{g[i]}) &= \sum_{g=1}^{G}(\beta_0 + \beta_1 \bar{X}_g + W_g)\Big(\sum_{i=1}^{n_g} X^{\perp}_{g[i]}\Big) \\
&= \sum_{g=1}^{G}(\beta_0 + \beta_1 \bar{X}_g + W_g)\Big(\sum_{i=1}^{n_g}(X_{g[i]} - \bar{X}_g)\Big) \\
&= \sum_{g=1}^{G}(\beta_0 + \beta_1 \bar{X}_g + W_g)\Big(n_g \bar{X}_g - n_g \bar{X}_g\Big) \\
&= 0 \qquad (36)
\end{aligned}$$

So, continuing Equation (35),

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{g,i} \epsilon_{g[i]}(X^{\perp}_{g[i]})}{\sum_{g,i}(X^{\perp}_{g[i]})^2} \tag{37}$$

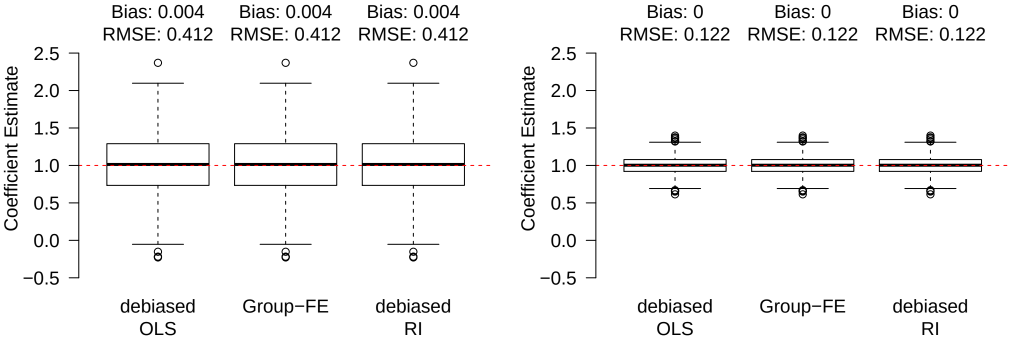
Because $\mathbb{E}(\epsilon_{g[i]} \mid X, Z) = 0$ in DGP 1, taking the expectation of the above yields $\mathbb{E}(\hat{\beta}_1) = \beta_1$. Finally, because $\bar{X}_g$ and $X^{\perp}_{g[i]}$ are uncorrelated (as $X^{\perp}_{g[i]}$ is mean-zero and $\sum_{g,i} \bar{X}_g(X^{\perp}_{g[i]}) = 0$ for the same reason as why Equation (36) simplifies to 0) and $\beta_1 X_{g[i]} + \alpha_1 \bar{X}_g$ can be rewritten as $\beta_1 X^{\perp}_{g[i]} + (\alpha_1 + \beta_1)\bar{X}_g$, this $\hat{\beta}_1$ is the the same as the estimate one would obtain from running an OLS including $\bar{X}_g$ as in Equation (13), proving the latter's unbiasedness.

□

A.6 Simulated example: adding group-level means may not debias $\hat{\beta}_{\text{MLM}}$ for group-level variables in the presence of correlated random effects

Consider again DGP 4 from Appendix A.4. As adding $\bar{U}^{(1)}_g = U^{(1)}_g$ to a model a second time is impossible, adding the group-level means of all included variables to a RI model will not eliminate the bias in estimating β_2. However, adding $\bar{X}^{(1)}_{g[i]}$ does debias estimates of β_1. See Figures 11 and 12 for the bias and RMSE of coefficient estimates in DGP 4 after adding $\bar{X}^{(1)}_{g[i]}$ to a RI model. For comparison, we also show the results from an OLS model that includes $\bar{X}^{(1)}_{g[i]}$ and a Group-FE model. However, the estimates from each model are exactly the same, with the exception of the nonexistent estimate of β_2 for Group-FE.

Figure 11. Comparison of estimates of β_1 from a RI model including $\bar{X}^{(1)}_g$, an OLS including $\bar{X}^{(1)}_g$, and Group-FE in DGP 4

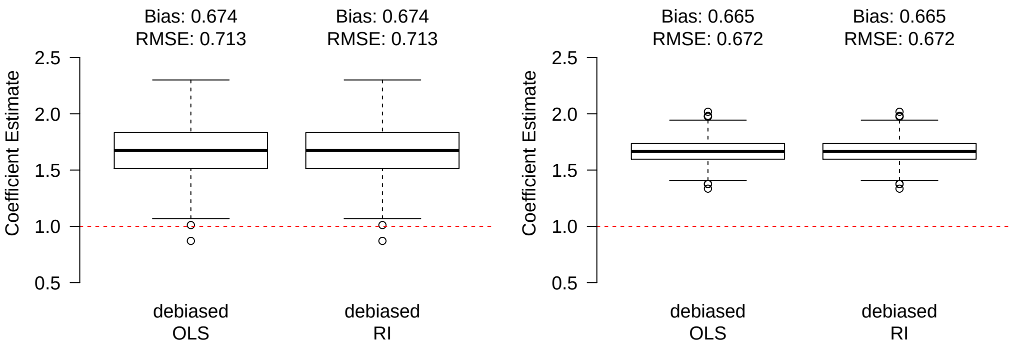


(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: The RI and OLS models have been debiased for β_1 by including $\bar{X}^{(1)}_g$ as a covariate. Results across 1000 iterations, each drawn from DGP 4 with $\beta_0 = \beta_1 = \beta_2 = 1$. The red dashed-line represents the true β_ℓ.

Figure 12. Comparison of estimates of β_2 from a RI model including $\bar{X}_g^{(1)}$ and an OLS including $\bar{X}_g^{(1)}$ in DGP 4



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: The RI and OLS models including $\bar{X}_g^{(1)}$ as a covariate are still referred to here as "debiased" because they unbiasedly estimate β_1, unlike RI and OLS models that omit $\bar{X}_g^{(1)}$, as can be seen in Figure 9. However, including $\bar{X}_g^{(1)}$ clearly does not debias their estimates of β_2. Results across 1000 iterations, each drawn from DGP 4 with $\beta_0 = \beta_1 = \beta_2 = 1$. The red dashed-line represents the true β_ℓ.

A.7 Simulated example: adding group-level means may induce bias in $\hat{\beta}_{\text{MLM}}$ for group-level variables when they are correlated with lower-level variables

Consider the following DGP:

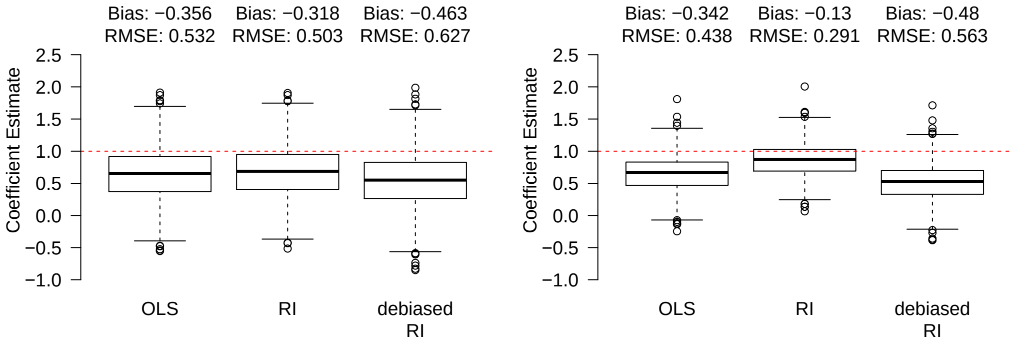
$$\begin{aligned} Y_{g[i]} &= \beta_0 + \beta_1 X_{g[i]}^{(1)} + \beta_2 U_g^{(1)} + (W_g^{(1)} + W_g^{(1)} W_g^{(2)}) + \epsilon_{g[i]} \qquad \text{(DGP 5)} \\ \text{where}\quad [W_g^{(1)}\ W_g^{(2)}]^\top &\overset{iid}{\sim} N(0, 2I_2), \\ X_{g[i]}^{(1)} &= W_g^{(1)} + W_g^{(2)} + N(0,1)_{g[i]} \\ U_g^{(1)} &= W_g^{(2)} + N(0,1)_g \\ \epsilon_{g[i]} &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

Here, there is an observed lower-level variable, $X_{g[i]}^{(1)}$, and an observed group-level variable, $U_g^{(1)}$. $X_{g[i]}^{(1)}$ is correlated with the unobserved random intercept, $(W_g^{(1)} + W_g^{(1)} W_g^{(2)})$, but even though $U_g^{(1)}$ is correlated with $W_g^{(2)}$, it is uncorrelated with the random intercept because it is independent of $W_g^{(1)}$. The inclusion of $\bar{X}_g^{(1)}$ in a RI model will therefore correct the bias in estimating β_1, and because $U_g^{(1)}$ is uncorrelated with the random intercept, one would imagine that such a model would also produce unbiased estimates of β_2. However, because $X_{g[i]}^{(1)}$ and $U_g^{(1)}$ are correlated, the inclusion of $\bar{X}_g^{(1)}$ in fact induces bias in estimates of β_2, which we see in Figure 13. All of the models show biases in estimates of β_2 at all sample sizes tried, with the RI model including $\bar{X}_g^{(1)}$ showing the most bias. However, unlike the estimates from an OLS and the RI model including $\bar{X}_g^{(1)}$, the estimates from the RI model *without* $\bar{X}_g^{(1)}$ here actually improve as n_g increases.

A.8 Proof of the unbiasedness of bcMLM with $\Sigma = \sigma^2 I_N$ under the conditional independence assumption

This proof has been adapted from Snijders and Berkhof 2008. For simplicity, we prove the result for when we have correctly assumed that $\beta_0 = 0$ in Equation (3), so the intercept term has been removed from $X_{g[i]}$, X_g, X, and β. Note that this does *not* prohibit an intercept term from being included in $Z_{g[i]}$.

Figure 13. Comparison of estimates of β_2 from an OLS, a RI model without $\bar{X}_g^{(1)}$, and an RI model including $\bar{X}_g^{(1)}$ in DGP 5



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: The RI model including $\bar{X}_g^{(1)}$ as a covariate is referred to here as "debiased" because it unbiasedly estimates β_1, whereas the RI model omitting $\bar{X}_g^{(1)}$ would not. However, including $\bar{X}_g^{(1)}$ clearly does not debias RI's estimate of β_2. Results across 1000 iterations, each drawn from DGP 5 with $\beta_0 = \beta_1 = \beta_2 = 1$. The red dashed-line represents the true β_ℓ.

First consider the projections of Y_g and X_g onto Z_g:

$$\tilde{Y}_g = Z_g(Z_g^\top Z_g)^{-1} Z_g^\top Y_g \tag{38}$$

$$\tilde{X}_g = Z_g(Z_g^\top Z_g)^{-1} Z_g^\top X_g \tag{39}$$

and then partialing out these projections from both Y_g and X_g, giving $Y_g^\perp$ and $X_g^\perp$:

$$\begin{aligned} Y_g^\perp &= Y_g - \tilde{Y}_g \\ &= [I_{n_g} - Z_g(Z_g^\top Z_g)^{-1} Z_g^\top] Y_g \end{aligned} \tag{40}$$

$$\begin{aligned} X_g^\perp &= X_g - \tilde{X}_g \\ &= [I_{n_g} - Z_g(Z_g^\top Z_g)^{-1} Z_g^\top] X_g \end{aligned} \tag{41}$$

Then let $\tilde{Y}$, $\tilde{X}$, $Y^\perp$, and $X^\perp$ be the (ordered) block matrices of the $\tilde{Y}_g$, $\tilde{X}_g$, $Y_g^\perp$, and $X_g^\perp$ respectively (as X is to the X_g). These matrices can also be written as

$$\tilde{Y} = Z(Z^\top Z)^{-1} Z^\top Y \tag{42}$$

$$\tilde{X} = Z(Z^\top Z)^{-1} Z^\top X \tag{43}$$

and

$$\begin{aligned} Y^\perp &= Y - \tilde{Y} \\ &= [I_N - Z(Z^\top Z)^{-1} Z^\top] Y \end{aligned} \tag{44}$$

$$\begin{aligned} X^\perp &= X - \tilde{X} \\ &= [I_N - Z(Z^\top Z)^{-1} Z^\top] X \end{aligned} \tag{45}$$

Then, consider an OLS predicting $Y^\perp$ with $X^\perp$. The resulting estimate of β would be

$$\hat{\beta} = [(X^\perp)^\top X^\perp]^{-1} (X^\perp)^\top Y^\perp \tag{46}$$

We will show that this $\hat{\beta}$ is unbiased, and is exactly equal to the estimate of β obtained from bcMLM.

Using that $Y = X\beta + Z\gamma + \epsilon$ and the definition of $Y^{\perp}$ in Equation (44), one finds

$$\begin{aligned}\hat{\beta} &= [(X^{\perp})^{\top}X^{\perp}]^{-1}(X^{\perp})^{\top}[I_N - Z(Z^{\top}Z)^{-1}Z^{\top}](X\beta + Z\gamma + \epsilon)\\ &= [(X^{\perp})^{\top}X^{\perp}]^{-1}(X^{\perp})^{\top}[I_N - Z(Z^{\top}Z)^{-1}Z^{\top}](X\beta + \epsilon)\end{aligned} \tag{47}$$

where the second equality in the above comes from the fact that $[I_N - Z(Z^{\top}Z)^{-1}Z^{\top}]Z = 0$. Then, using the definition of $X^{\perp}$ in Equation (45) and that $[I_N - Z(Z^{\top}Z)^{-1}Z^{\top}]$ is idempotent, the above becomes

$$\begin{aligned}\hat{\beta} &= [(X^{\perp})^{\top}X^{\perp}]^{-1}[(X^{\perp})^{\top}X^{\perp}]\beta + [(X^{\perp})^{\top}X^{\perp}]^{-1}(X^{\perp})^{\top}\epsilon\\ &= \beta + [(X^{\perp})^{\top}X^{\perp}]^{-1}(X^{\perp})^{\top}\epsilon\end{aligned} \tag{48}$$

which means that $\mathbb{E}(\hat{\beta}) = \beta$, because $\mathbb{E}(\epsilon \mid X, Z) = 0$ by the conditional independence assumption. So, we have shown that an OLS of $Y^{\perp}$ on $X^{\perp}$ yields an unbiased estimate of β.

Now we consider the optimization problem solved in bcMLM to find an estimate of β, and show it yields the unbiased $\hat{\beta}$ above. bcMLM with $\Sigma = \sigma^2 I_N$ operates under the assumption that $Y_g \mid X, Z \overset{iid}{\sim} \mathcal{N}(X_g\beta + \tilde{X}_g\alpha, V_g)$ where $V_g = Z_g\Omega Z_g^{\top} + \sigma^2 I_{n_g}$. This implies that contribution of the g^{th} group to the assumed log likelihood for the model is:

$$\ell_g^{(\text{bcMLM})} = -\frac{1}{2}\log|V_g| - \frac{1}{2}[Y_g - X_g\beta - \tilde{X}_g\alpha]^{\top}V_g^{-1}[Y_g - X_g\beta - \tilde{X}_g\alpha] \tag{49}$$

Using that

$$\begin{aligned} & V_g^{-1} = \sigma^{-2}I_{n_g} - Z_g A_g Z_g^{\top}\\ \text{where } \; & A_g = \sigma^{-2}(Z_g^{\top}Z_g)^{-1} - (Z_g^{\top}Z_g)^{-1}[\sigma^2(Z_g^{\top}Z_g)^{-1} + \Omega]^{-1}(Z_g^{\top}Z_g)^{-1}\end{aligned} \tag{50}$$

and that $Y_g - X_g\beta - \tilde{X}_g\alpha = [Y_g^{\perp} - X_g^{\perp}\beta] + [\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]$, it can be shown that

$$\ell_g^{(\text{bcMLM})} = -\frac{1}{2}\left(\log|V_g| + \sigma^{-2}||Y_g^{\perp} - X_g^{\perp}\beta||_2^2 + [\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]^{\top}V_g^{-1}[\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]\right) \tag{51}$$

So, the whole likelihood maximization procedure, given Ω and σ^2 (which one estimates first in bcMLM before finding estimates of β and α) is

$$\begin{aligned} & \operatorname*{arg\,max}_{\alpha,\beta} \sum_{g=1}^{G} \ell_g^{(\text{bcMLM})}\\ &= \operatorname*{arg\,max}_{\alpha,\beta} -\frac{1}{2}\sum_{g=1}^{G}\left(\log|V_g| + \sigma^{-2}||Y_g^{\perp} - X_g^{\perp}\beta||_2^2 + [\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]^{\top}V_g^{-1}[\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]\right)\\ &= \operatorname*{arg\,min}_{\alpha,\beta} \sum_{g=1}^{G}\left(||Y_g^{\perp} - X_g^{\perp}\beta||_2^2 + \sigma^2[\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]^{\top}V_g^{-1}[\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]\right)\\ &= \operatorname*{arg\,min}_{\alpha,\beta} \left(||Y^{\perp} - X^{\perp}\beta||_2^2 + \sigma^2\sum_{g=1}^{G}[\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]^{\top}V_g^{-1}[\tilde{Y}_g - \tilde{X}_g(\beta + \alpha)]\right)\end{aligned} \tag{52}$$

We see in the above minimization problem that when choosing β, the only part of the objective function that matters is $||Y^{\perp} - X^{\perp}\beta||_2^2$, which is the same objective function as that in an OLS predicting $Y^{\perp}$ with $X^{\perp}$. This means that the estimate of β from bcMLM here is exactly the $\hat{\beta}$ defined earlier in the proof (Equation (46)), which we have shown is unbiased. Therefore, bcMLM with $\Sigma = \sigma^2 I_N$ will produce an unbiased estimate of β.

□

A.9 Proof of the equivalance of FE and bcMLM with $\Sigma = \sigma^2 I_N$

Another consequence of the proof of the unbiasedness of bcMLM with $\Sigma = \sigma^2 I_N$ provided in Appendix A.8 is that there is an exact equivalence between the estimates of β from bcMLM and FE. This is because, like bcMLM, FE produces the same estimate of β as does an OLS regression of $Y^\perp$ on $X^\perp$ (defined in Appendix A.8), as FE can be reparametrized as

$$\begin{aligned} Y &= X\beta + Z\gamma + \epsilon \\ &= [I_N - Z(Z^\top Z)^{-1}Z^\top]X\beta + [Z(Z^\top Z)^{-1}Z^\top]X\beta + Z\gamma + \epsilon \\ &= X^\perp\beta + Z\tilde{\gamma} + \epsilon \end{aligned} \tag{53}$$

where $\tilde{\gamma} = (Z^\top Z)^{-1}Z^\top X\beta + \gamma$. Because $X^\perp$ and Z are orthogonal (i.e., $Z^\top X^\perp = 0$), the estimate of β obtained by FE is therefore

$$\begin{aligned} \hat{\beta}_{\text{FE}} &= [(X^\perp)^\top X^\perp]^{-1}(X^\perp)^\top Y \\ &= [(X^\perp)^\top X^\perp]^{-1}X^\top[I_N - Z(Z^\top Z)^{-1}Z^\top]Y \\ &= [(X^\perp)^\top X^\perp]^{-1}X^\top[I_N - Z(Z^\top Z)^{-1}Z^\top]^2 Y \\ &= [(X^\perp)^\top X^\perp]^{-1}(X^\perp)^\top Y^\perp \end{aligned} \tag{54}$$

where we have used in the third line of Equation (54) above that $[I_N - Z(Z^\top Z)^{-1}Z^\top]$ is idempotent.

□

A.10 Simulated example: biased $\hat{\beta}_{\text{MLM}}$ with a random slope

While we are primarily concerned with the RI specification of MLM, we also consider here when MLM produces biased $\hat{\beta}_{\text{MLM}}$ due to a random *slope*. Consider the DGP

$$Y_{g[i]} = \beta_0 + (\beta_1 + W_g^{(1)})X_{g[i]}^{(1)} + \epsilon_{g[i]} \tag{55}$$

where $W_g^{(1)}$ is an unobserved group-level variable. If the above were fit by a simple OLS of Y on X or MLM, we would expect a biased estimate of β_1 when $\text{cov}(X_{g[i]}^{(1)}, W_g^{(1)}X_{g[i]}^{(1)}) \neq 0$. Assuming $X_{g[i]}^{(1)}$ and $W_g^{(1)}$ are both mean-zero, this occurs when

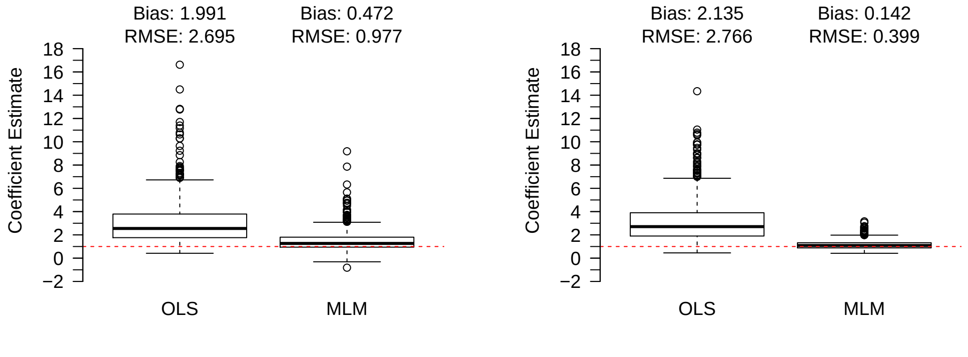
$$\begin{aligned} \text{cov}(X_{g[i]}^{(1)}, W_g^{(1)}X_{g[i]}^{(1)}) &= \mathbb{E}\Big([X_{g[i]}^{(1)}]^2 W_g^{(1)}\Big) - \mathbb{E}(X_{g[i]}^{(1)})\mathbb{E}(W_g^{(1)}X_{g[i]}^{(1)}) \\ &= \mathbb{E}\Big[W_g^{(1)}\mathbb{E}\Big([X_{g[i]}^{(1)}]^2 \mid W_g^{(1)}\Big)\Big] \\ &\neq 0 \end{aligned} \tag{56}$$

There are many ways for the above to hold, but if $X_{g[i]} = W_g^{(1)}N(0,1)_{g[i]} + N(0,1)_{g[i]}$, then the above would require $\mathbb{E}([W_g^{(1)}]^3) \neq 0$, i.e., $W_g^{(1)}$ has an asymmetric distribution. We consider such a case in DGP 6 below:

$$\begin{aligned} & Y_{g[i]} = \beta_0 + (\beta_1 + W_g^{(1)})X_{g[i]}^{(1)} + \epsilon_{g[i]} \\ \text{where} \quad & W_g^{(1)} \overset{iid}{\sim} \mathcal{X}_1^2 - 1 \\ & X_{g[i]}^{(1)} = W_g^{(1)}N(0,1)_{g[i]} + N(0,1)_{g[i]} \\ & \epsilon_{g[i]} \overset{iid}{\sim} N(0,\sigma^2) \end{aligned} \tag{DGP 6}$$

In DGP 6, $W_g^{(1)}$ is a centered chi-squared with one degree of freedom, and thus satisfies $\mathbb{E}([W_g^{(1)}]^3) \neq 0$. We see biases in estimates of β_1 from a simple OLS and a MLM with a random intercept and slope for $X_{g[i]}^{(1)}$ across draws from DGP 6 with $\beta_0 = \beta_1 = 1$ in Figure 14.

Figure 14. Comparison of estimates of β_1 from OLS and a MLM with a random intercept and slope for $X_{g[i]}^{(1)}$ in DGP 6



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: Results across 1000 iterations, each drawn from DGP 6 with $\beta_0 = \beta_1 = 1$. The red dashed-line represents the true β_ℓ.

A.11 Simulated example: including $(\bar{X}_g - \bar{X})^\top \otimes Z_{g[i]}$ in MLM does not alleviate bias from DGP 6

We consider here the proposed solution to correlated random effects of including $(\bar{X}_g - \bar{X})^\top \otimes Z_{g[i]}$ as fixed effect variables presented in Snijders and Bosker 2011 and Wooldridge 2013.

In DGP 6 from Appendix A.10, this proposal would imply the MLM

$$\begin{aligned} Y_{g[i]} &= \beta_0 + \beta_1 X_{g[i]}^{(1)} + \alpha_0(\bar{X}_g^{(1)} - \bar{X}^{(1)}) + \alpha_1(\bar{X}_g^{(1)} - \bar{X}^{(1)})X_{g[i]}^{(1)} \\ &\quad + \gamma_{0g} + \gamma_{1g} X_{g[i]}^{(1)} + \epsilon_{g[i]} \\ \text{where} \quad & [\gamma_{0g}\ \gamma_{1g}]^\top \overset{iid}{\sim} \mathcal{N}(0, \Omega) \end{aligned} \tag{57}$$

and $\epsilon_{g[i]} \overset{iid}{\sim} N(0, \sigma^2)$. However, the above model does not correct the bias shown by a MLM with a random intercept and slope in DGP 6, as we see Figure 15. Estimates from a per-cluster regression (introduced by Bates *et al.* 2014, and described in Appendix A.13), on the other hand, are unbiased.

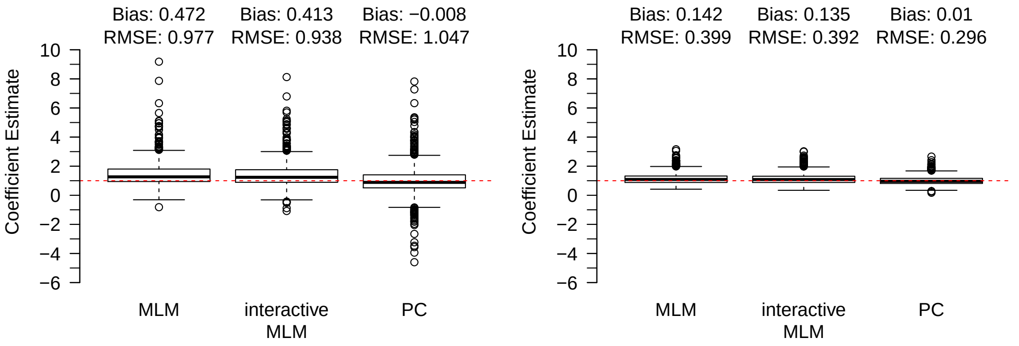
A.12 Simulated example: per-cluster regression to unbiasedly estimate coefficients of group-level variables that are uncorrelated with random effects

Consider again DPG 5 introduced in Appendix A.7. After estimating β_1 with an unbiased $\hat{\beta}_1$ obtained by FE or adding the group-mean of $X_{g[i]}^{(1)}$ to a RI model, one can apply the per-cluster regression to unbiasedly estimate β_2 because $U_g^{(1)}$ is uncorrelated with the random intercept, $(W_g^{(1)} + W_g^{(1)} W_g^{(2)})$.

The first step in the per-cluster regression is to subtract the estimated marginal effect of $X_{g[i]}^{(1)}$ from $Y_{g[i]}$, like so:

$$Y_{g[i]}^{\perp} = Y_{g[i]} - \hat{\beta}_1 X_{g[i]}^{(1)} \qquad \text{(Per-cluster Regression: Step 1)}$$

Figure 15. Comparison of estimates of β_1 from a standard MLM, the MLM in Equation (57), and a per-cluster regression in DGP 6



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: "PC" refers to the per-cluster regression, and the MLM in Equation (57) is referred to as "interactive" because it includes the interaction of $(\bar{X}_g - \bar{X})$ with all variables in $Z_{g[i]}$. Results across 1000 iterations, each drawn from DGP 6 with $\beta_0 = \beta_1 = 1$. The red dashed-line represents the true β_ℓ.

The next step is to calculate the group-means of $Y^{\perp}_{g[i]}$:

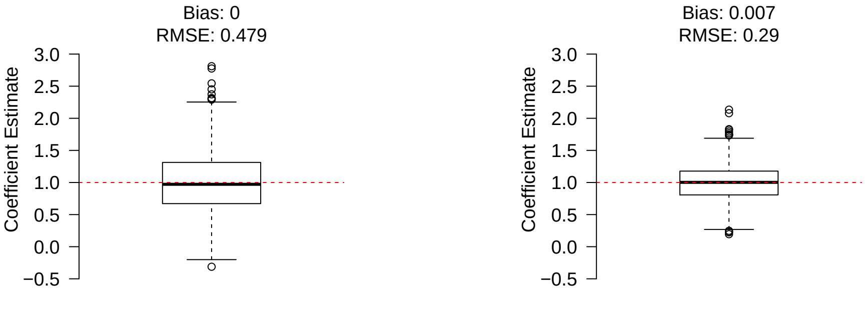
$$\hat{\eta}_{0g} = \frac{1}{n_g}\sum_{i=1}^{n_g} Y^{\perp}_{g[i]} \qquad \text{(Per-cluster Regression: Step 2)}$$

Finally, regressing $\hat{\eta}_{0g}$ on $U^{(1)}_g$ and an intercept term in a group-level OLS provides an unbiased estimate of β_2:

$$\text{Estimate } \hat{\eta}_{0g} = \beta_0 + \beta_2 U^{(1)}_g + \delta_{0g} \text{ by OLS} \qquad \text{(Per-cluster Regression: Step 3)}$$

where δ_{0g} acts as the residual. In Figure 16, we see that the estimate of β_2 from a per-cluster regression is unbiased in DGP 5, unlike that from OLS, and RI including or without $\bar{X}^{(1)}_g$.

Figure 16. Estimates of β_2 from a per-cluster regression in DGP 5



(a) $G = 50,\ n_g = 5$

(b) $G = 50,\ n_g = 50$

Note: Results across 1000 iterations, each drawn from DGP 5 with $\beta_0 = \beta_1 = \beta_2 = 1$. The red dashed-line represents the true β_ℓ.

A.13 Per-cluster regression for varying slopes and cross-level interactions

As noted in the text, another example of over-identification in bcMLM is when the slope for $X_{g[i]}^{(\ell)}$ is allowed to vary, i.e., $X_{g[i]}^{(\ell)}$ is included in $Z_{g[i]}$. Because bcMLM includes as extra covariates the predictions of $X_{g[i]}^{(\ell)}$ using $Z_{g[i]}$, and $Z_{g[i]}$ predicts $X_{g[i]}^{(\ell)}$ perfectly, including this "prediction" as an extra variable simply includes $X_{g[i]}^{(\ell)}$ in the model twice. Therefore, one of the $X_{g[i]}^{(\ell)}$ will be dropped out of the model, and the "prediction" of $X_{g[i]}^{(\ell)}$ 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, $X_{g[i]}^{(\ell)}U_g^{(k)}$.

The per-cluster regression provides an option for users who are interested in those coefficients. Let $X_{g[i]}^{(sub)}$ be the sub-vector of $X_{g[i]}$ containing the variables that are *not* predicted perfectly by (i.e., colinear with) $Z_{g[i]}$ (e.g., neither the variables in both $X_{g[i]}$ and $Z_{g[i]}$ nor their cross-level interactions included in $X_{g[i]}$) with corresponding coefficients $\beta^{(sub)}$, and let $X_g^{(sub)}$ be the corresponding sub-matrix of X_g. Furthermore, let $X_{g[i]}^{(\ell)}$, the $(\ell+1)$th element of $X_{g[i]}$ with coefficient β_ℓ, also be the $(\ell+1)$th element $Z_{g[i]}$ (i.e., $Z_{g[i]}^{(\ell)} = X_{g[i]}^{(\ell)}$), and let $(X_{g[i]}^{(\ell)}U_g^{(1)}, \dots, X_{g[i]}^{(\ell)}U_g^{(r)})$ be the r cross-level interactions of $X_{g[i]}^{(\ell)}$ included in $X_{g[i]}$ with corresponding coefficients $(\beta_{\ell+1}, \dots, \beta_{\ell+r})$.

A per-cluster regression approach proceeds by first unbiasedly estimating $\beta^{(sub)}$, using bcMLM or FE. The estimated marginal effects of $X_g^{(sub)}$ must then be purged from Y_g, forming $Y_g^{\perp} = Y_g - X_g^{(sub)}\hat{\beta}^{(sub)}$. Next, regress each of the G vectors $Y_g^{\perp}$ on Z_g individually by OLS, obtaining G coefficient vectors $\hat{\eta}_g = (Z_g^\top Z_g)^{-1}Z_g^\top Y_g^{\perp}$. Finally, letting $\hat{\eta}_{\ell g}$ be the $(\ell+1)$th element of each $\hat{\eta}_g$ (i.e., the coefficient for $Z_{g[i]}^{(\ell)}$ from the OLS in the previous step), regress $\hat{\eta}_{\ell g}$ on an intercept term and $(U_g^{(1)}, \dots, U_g^{(r)})$ in a group-level regression fit by OLS, as in an assumed model:

$$\hat{\eta}_{\ell g} = \beta_\ell + \sum_{k=1}^{r} \beta_{\ell+k}U_g^{(k)} + \delta_{\ell g} \tag{58}$$

where $\delta_{\ell g}$ acts as the residual. If $\mathbb{E}(\gamma_{\ell g} \mid U_g^{(1)}, \dots, U_g^{(r)}) = 0$, the estimated coefficients from this final step are unbiased for $(\beta_\ell, \dots, \beta_{\ell+r})$. Note that in the case where there are no cross-level interactions with $X_{g[i]}^{(\ell)}$, this final step amounts to simply taking the mean of $\hat{\eta}_{\ell g}$ over the G groups as the estimate of β_ℓ, and the condition for its unbiasedness is that the $\gamma_{\ell g}$ are unconditionally mean-zero. Furthermore, note that if $X_{g[i]}^{(\ell)} = 1$, then $(\beta_{\ell+1}, \dots, \beta_{\ell+r})$ are the coefficients for the group-level variables, $(U_g^{(1)}, \dots, U_g^{(r)})$.

A.14 Discussion of cluster-robust standard errors

To see why CRSEs work in theory, consider forming them after an OLS of Y on X, where $Y_g = X_g\beta + \epsilon_g^*$ and the $\epsilon_g^* \in \mathbb{R}^{n_g}$ are mutually independent and mean-zero given X, but have unknown covariance matrices, $\mathrm{var}(\epsilon_g^* \mid X) = \mathbb{E}(\epsilon_g^*\epsilon_g^{*\top} \mid X)$. The OLS estimate of β is $\hat{\beta}_{\mathrm{OLS}} = (X^\top X)^{-1}X^\top Y$, meaning

$$\begin{aligned}\mathrm{var}(\hat{\beta}_{\mathrm{OLS}} \mid X) &= (X^\top X)^{-1}X^\top \mathrm{var}(Y \mid X)X(X^\top X)^{-1} \\ &= (X^\top X)^{-1}X^\top \begin{bmatrix} \mathbb{E}(\epsilon_1^*\epsilon_1^{*\top} \mid X) & & 0 \\ & \ddots & \\ 0 & & \mathbb{E}(\epsilon_G^*\epsilon_G^{*\top} \mid X) \end{bmatrix} X(X^\top X)^{-1} \end{aligned} \tag{59}$$

That CRSEs use $\hat{\mathbb{E}}_{\text{CRSE}}(\epsilon^*_{g[i]}\epsilon^*_{g[i']} \mid X) = c \times \hat{e}_{g[i]}\hat{e}_{g[i']}$ where $\hat{e}_{g[i]} = Y_{g[i]} - X^\top_{g[i]}\hat{\beta}_{\text{OLS}}$ implies

$$\hat{\mathbb{E}}_{\text{CRSE}}(\epsilon^*_g\epsilon^{*\top}_g \mid X) = c \times \begin{bmatrix} \hat{e}^2_{g[1]} & \hat{e}_{g[1]}\hat{e}_{g[2]} & \cdots & \hat{e}_{g[1]}\hat{e}_{g[n_g]} \\ \hat{e}_{g[2]}\hat{e}_{g[1]} & \hat{e}^2_{g[2]} & \cdots & \hat{e}_{g[2]}\hat{e}_{g[n_g]} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}_{g[n_g]}\hat{e}_{g[1]} & \hat{e}_{g[n_g]}\hat{e}_{g[2]} & \cdots & \hat{e}^2_{g[n_g]} \end{bmatrix}$$
$$= c \times \hat{e}_g\hat{e}^\top_g \tag{60}$$

where $\hat{e}_g = Y_g - X_g\hat{\beta}_{\text{OLS}}$. Therefore, $\widehat{\text{var}}_{\text{CRSE}}(\hat{\beta}_{\text{OLS}})$ is

$$\widehat{\text{var}}_{\text{CRSE}}(\hat{\beta}_{\text{OLS}}) = c \times (X^\top X)^{-1}X^\top \begin{bmatrix} \hat{e}_1\hat{e}^\top_1 & & 0 \\ & \ddots & \\ 0 & & \hat{e}_G\hat{e}^\top_G \end{bmatrix} X(X^\top X)^{-1} \tag{61}$$

Remembering that $X = \begin{bmatrix} X_1 \\ \vdots \\ X_G \end{bmatrix}$ allows one to rewrite this as

$$\widehat{\text{var}}_{\text{CRSE}}(\hat{\beta}_{\text{OLS}}) = c \times (\sum_{g=1}^{G} X^\top_g X_g)^{-1}(\sum_{g=1}^{G} X^\top_g \hat{e}_g\hat{e}^\top_g X_g)(\sum_{g=1}^{G} X^\top_g X_g)^{-1}$$
$$= \frac{c}{G} \times (\frac{1}{G}\sum_{g=1}^{G} X^\top_g X_g)^{-1}(\frac{1}{G}\sum_{g=1}^{G} X^\top_g \hat{e}_g\hat{e}^\top_g X_g)(\frac{1}{G}\sum_{g=1}^{G} X^\top_g X_g)^{-1} \tag{62}$$

and the averaging over the G groups in each of the summations above is why CRSEs can “learn" any dependence structure when G is sufficiently large.[29]

A.15 Choice of c for CRSEs with MLM and bcMLM

In choosing c, Cameron and Miller 2015 write that the common choice for a simple OLS of Y on X is $c = \frac{G}{G-1}\frac{N-1}{N-p}$. This is the typical c employed in MLM as well, but the decision is more complicated for bcMLM. In the case of solely varying intercepts, it would be tempting to employ $c = \frac{G}{G-1}\frac{N-1}{N-(p+\bar{p})}$ in bcMLM, where $\bar{p}$ is the number group-level means that have been added. However, since there is an exact equivalence between estimates of β from Group-FE and bias-corrected RI when $\Sigma = \sigma^2 I_N$, we suggest using $c = \frac{G}{G-1}\frac{N-1}{N-(p+G-1)}$. This is the common choice for Group-FE (Cameron and Miller 2015), as it accounts for the extra $G - 1$ group indicator variables included in the model. For a general bcMLM, we make a similar recommendation. Every extra group-varying slope in a FE model requires an extra $G - 1$ variables, as $G - 1$ group indicator variables are interacted with the variable whose slope is to vary. Therefore, an FE model that allows d varying coefficients would require $c = \frac{G}{G-1}\frac{N-1}{N-[p+d(G-1)]}$. For a general bcMLM with d varying slopes or intercepts, we also suggest this choice of c.

A.16 Proof of the equivalance of CRSEs from bcMLM with $\Sigma = \sigma^2 I_N$ and FE

Like in Appendix A.8, we prove this for the case where intercept term has been removed from $X_{g[i]}$, X_g, X, and β, meaning $\beta = (\beta_1, \ldots, \beta_{p-1})$. Additionally, we largely here use the notation from

[29] As mentioned in Section 3.3, $G = 50$ is generally viewed as the large enough (Cameron and Miller 2015). However, there is far from a consensus, and the true benchmark will differ by the situation.

Appendix A.8, but review it for convenience.

Let $[(X_g - \tilde{X}_g)\ \tilde{X}_g]$ be the matrix, for group g, of variables included as fixed effect variables in an equivalent form of bcMLM where X_g has been centered by $\tilde{X}_g = Z_g(Z_g^\top Z_g)^{-1} Z_g^\top X_g$. Then, let $\tilde{X}$ be the stacked matrix of $\tilde{X}_g$ (as X is to X_g), and let (β, α) be the coefficient vector for $[(X_g - \tilde{X}_g)\ \tilde{X}_g]$ to be estimated. $\tilde{X}$ can also be expressed as $\tilde{X} = Z(Z^\top Z)^{-1} Z^\top X$. Furthermore, let the bcMLM estimate of $\operatorname{var}(Y \mid X, Z) = V$ be

$$\begin{aligned} \hat{V} &= Z \begin{bmatrix} \hat{\Omega} & & 0 \\ & \ddots & \\ 0 & & \hat{\Omega} \end{bmatrix} Z^\top + \hat{\sigma}^2 I_N \\ &= \begin{bmatrix} Z_1 \hat{\Omega} Z_1^\top + \hat{\sigma}^2 I_{n_1} & & 0 \\ & \ddots & \\ 0 & & Z_G \hat{\Omega} Z_G^\top + \hat{\sigma}^2 I_{n_G} \end{bmatrix} \\ &= \begin{bmatrix} \hat{V}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{V}_G \end{bmatrix} \end{aligned} \tag{63}$$

where $\hat{V}_g = Z_g \hat{\Omega} Z_g^\top + \hat{\sigma}^2 I_{n_g}$ and $(\hat{\Omega}, \hat{\sigma}^2)$ is the bcMLM estimate of (Ω, σ^2).

Similarly, we consider the alternate, but isomorphic, form of FE where X_g has been centered by $\tilde{X}_g$. Note that this form of FE is equivalent to that introduced in Section 2.2 because:

$$\begin{aligned} (X_{g[i]} - \tilde{X}_{g[i]}^\top)^\top \beta + Z_{g[i]}^\top \gamma_g &= X_{g[i]}^\top \beta - Z_{g[i]}^\top (Z_g^\top Z_g)^{-1} Z_g^\top X_g \beta + Z_{g[i]}^\top \gamma_g \\ &= X_{g[i]}^\top \beta - Z_{g[i]}^\top \tilde{\gamma}_g \end{aligned} \tag{64}$$

where $\tilde{\gamma}_g = (Z_g^\top Z_g)^{-1} Z_g^\top X_g \beta + \gamma_g$. Note that we consider the $\tilde{X}_{g[i]}$-centered versions of bcMLM and FE because $X_g - \tilde{X}_g$ is orthogonal to Z_g and $\tilde{X}_g$ (i.e., $(X_g - \tilde{X}_g)^\top Z_g = (X_g - \tilde{X}_g)^\top \tilde{X}_g = 0$), and likewise $X - \tilde{X}$ is orthogonal to Z and $\tilde{X}$. This greatly simplifies the matrix algebra.

Finally, let $\hat{\beta}$ be the estimate of β that comes from bcMLM and FE (as a reminder, they are equivalent), let $\hat{\gamma}_g$ be the estimate of γ_g from FE, and let $\hat{\alpha}$ be the estimate of α from bcMLM. With this notation, the residual vectors for group g from each model are

$$\hat{e}_g^{(\text{bcMLM})} = Y_g - (X_g - \tilde{X}_g)\hat{\beta} - \tilde{X}_g \hat{\alpha} \tag{bcMLM Residual}$$

$$\hat{e}_g^{(\text{FE})} = Y_g - (X_g - \tilde{X}_g)\hat{\beta} - Z_g \hat{\gamma}_g \tag{FE Residual}$$

For FE, the CRSE estimator of the variance of $\hat{\beta}$ is the first $p - 1$ rows and columns of

$$c \times B^{(\text{FE})} \times M^{(\text{FE})} \times B^{(\text{FE})} \tag{65}$$

where

$$\begin{aligned}
B^{(\mathrm{FE})} &= \left(\begin{bmatrix} (X - \tilde{X})^\top \\ Z^\top \end{bmatrix} \begin{bmatrix} (X - \tilde{X}) & Z \end{bmatrix} \right)^{-1} \\
&= \begin{bmatrix} \sum_{g=1}^{G} (X_g - \tilde{X}_g)^\top (X_g - \tilde{X}_g) & (X - \tilde{X})^\top Z \\ Z^\top (X - \tilde{X}) & Z^\top Z \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \left(\sum_{g=1}^{G} (X_g - \tilde{X}_g)^\top (X_g - \tilde{X}_g) \right)^{-1} & 0 \\ 0 & (Z^\top Z)^{-1} \end{bmatrix} \qquad (66)
\end{aligned}$$

and

$$M^{(\mathrm{FE})} = \begin{bmatrix} (X - \tilde{X})^\top \\ Z^\top \end{bmatrix} \begin{bmatrix} \hat{e}_1^{(\mathrm{FE})} [\hat{e}_1^{(\mathrm{FE})}]^\top & & 0 \\ & \ddots & \\ 0 & & \hat{e}_G^{(\mathrm{FE})} [\hat{e}_G^{(\mathrm{FE})}]^\top \end{bmatrix} \begin{bmatrix} (X - \tilde{X}) & Z \end{bmatrix} \qquad (67)$$

Given Equation (65), and because $B^{(\mathrm{FE})}$ is block diagonal, the only portion of $M^{(\mathrm{FE})}$ that contributes to the variance of $\hat{\beta}$ is its first $p - 1$ rows and columns, i.e.,

$$\begin{aligned}
&(X - \tilde{X})^\top \begin{bmatrix} \hat{e}_1^{(\mathrm{FE})} [\hat{e}_1^{(\mathrm{FE})}]^\top & & 0 \\ & \ddots & \\ 0 & & \hat{e}_G^{(\mathrm{FE})} [\hat{e}_G^{(\mathrm{FE})}]^\top \end{bmatrix} (X - \tilde{X}) \\
&\quad = \sum_{g=1}^{G} (X_g - \tilde{X}_g)^\top \hat{e}_g^{(\mathrm{FE})} [\hat{e}_g^{(\mathrm{FE})}]^\top (X_g - \tilde{X}_g) \\
&\quad = \sum_{g=1}^{G} \left((X_g - \tilde{X}_g)^\top \hat{e}_g^{(\mathrm{FE})} \right) \left((X_g - \tilde{X}_g)^\top \hat{e}_g^{(\mathrm{FE})} \right)^\top \qquad (68)
\end{aligned}$$

Expanding $(X_g - \tilde{X}_g)^\top \hat{e}_g^{(\mathrm{FE})}$ in Equation (68) yields

$$\begin{aligned}
(X_g - \tilde{X}_g)^\top \hat{e}_g^{(\mathrm{FE})} &= (X_g - \tilde{X}_g)^\top [Y_g - (X_g - \tilde{X}_g)\hat{\beta} - Z_g \hat{\gamma}_g] \\
&= (X_g - \tilde{X}_g)^\top (Y_g - X_g \hat{\beta}) \qquad (69)
\end{aligned}$$

which substituting into Equation (68) yields

$$\sum_{g=1}^{G} (X_g - \tilde{X}_g)^\top (Y_g - X_g \hat{\beta})(Y_g - X_g \hat{\beta})^\top (X_g - \tilde{X}_g) \qquad (70)$$

Therefore, as Equation (70) is the first $p - 1$ rows and columns of $M^{(\text{FE})}$, the CRSEs for $\hat{\beta}$ from FE are

$$\begin{aligned}\widehat{\text{var}}^{(\text{FE})}_{\text{CRSE}}(\hat{\beta}) = c\times & \Big(\sum_{g=1}^{G}(X_g - \tilde{X}_g)^{\top}(X_g - \tilde{X}_g)\Big)^{-1} \\ & \times\Big(\sum_{g=1}^{G}(X_g - \tilde{X}_g)^{\top}(Y_g - X_g\hat{\beta})(Y_g - X_g\hat{\beta})^{\top}(X_g - \tilde{X}_g)\Big) \\ & \times\Big(\sum_{g=1}^{G}(X_g - \tilde{X}_g)^{\top}(X_g - \tilde{X}_g)\Big)^{-1} \end{aligned} \tag{71}$$

Turning to bcMLM, its CRSE variance estimator for $\hat{\beta}$ is the first $p - 1$ rows and columns of

$$c \times B^{(\text{bcMLM})} \times M^{(\text{bcMLM})} \times B^{(\text{bcMLM})} \tag{72}$$

where

$$\begin{aligned} B^{(\text{bcMLM})} &= \Bigg(\begin{bmatrix}(X - \tilde{X})^{\top} \\ \tilde{X}^{\top}\end{bmatrix}\hat{V}^{-1}\begin{bmatrix}(X - \tilde{X}) & \tilde{X}\end{bmatrix}\Bigg)^{-1} \\ &= \begin{bmatrix}\sum_{g=1}^{G}(X_g - \tilde{X}_g)^{\top}\hat{V}_g^{-1}(X_g - \tilde{X}_g) & \sum_{g=1}^{G}(X_g - \tilde{X}_g)^{\top}\hat{V}_g^{-1}\tilde{X}_g \\ \sum_{g=1}^{G}\tilde{X}_g^{\top}\hat{V}_g^{-1}(X_g - \tilde{X}_g) & \sum_{g=1}^{G}\tilde{X}_g^{\top}\hat{V}_g^{-1}\tilde{X}_g\end{bmatrix}^{-1} \end{aligned} \tag{73}$$

and

$$\begin{aligned} M^{(\text{bcMLM})} = & \begin{bmatrix}(X - \tilde{X})^{\top} \\ \tilde{X}^{\top}\end{bmatrix}\hat{V}^{-1} \\ & \times\begin{bmatrix}\hat{e}_1^{(\text{bcMLM})}[\hat{e}_1^{(\text{bcMLM})}]^{\top} & & 0 \\ & \ddots & \\ 0 & & \hat{e}_G^{(\text{bcMLM})}[\hat{e}_G^{(\text{bcMLM})}]^{\top}\end{bmatrix} \\ & \times\hat{V}^{-1}\begin{bmatrix}(X - \tilde{X}) & \tilde{X}\end{bmatrix} \end{aligned} \tag{74}$$

Starting with $B^{(\text{bcMLM})}$, we first show that the off-diagonal matrices are 0. Like in Appendix A.8, we use the fact that

$$\begin{aligned} & V_g^{-1} = \sigma^{-2}I_{n_g} - Z_g A_g Z_g^{\top} \\ \text{where} \quad & A_g = \sigma^{-2}(Z_g^{\top}Z_g)^{-1} - (Z_g^{\top}Z_g)^{-1}[\sigma^2(Z_g^{\top}Z_g)^{-1} + \Omega]^{-1}(Z_g^{\top}Z_g)^{-1} \end{aligned} \tag{75}$$

Therefore, the off-diagonal matrices in $B^{(\text{bcMLM})}$ given in Equation (73) are

$$\begin{aligned} \sum_{g=1}^{G}\tilde{X}_g^{\top}\hat{V}_g^{-1}(X_g - \tilde{X}_g) &= \sum_{g=1}^{G}\tilde{X}_g^{\top}(\hat{\sigma}^{-2}I_{n_g} - Z_g\hat{A}_g Z_g^{\top})(X_g - \tilde{X}_g) \\ &= 0 \end{aligned} \tag{76}$$

Moving to the first $p - 1$ rows and columns of $B^{(\text{bcMLM})}$ given in Equation (73), we can again use the expression for V_g^{-1} in Equation (75) to obtain

$$\begin{aligned}\sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top \hat{V}_g^{-1}(X_g - \tilde{X}_g) &= \sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top(\hat{\sigma}^{-2}I_{n_g} - Z_g\hat{A}_g Z_g^\top)(X_g - \tilde{X}_g) \\ &= \hat{\sigma}^{-2}\sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top(X_g - \tilde{X}_g) \qquad (77)\end{aligned}$$

Substituting Equations (76) and (77) into Equation (73), we find

$$\begin{aligned}B^{(\mathrm{bcMLM})} &= \begin{bmatrix}\hat{\sigma}^{-2}\sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top(X_g - \tilde{X}_g) & 0 \\ 0 & \sum_{g=1}^{G}\tilde{X}_g^\top V_g^{-1}\tilde{X}_g\end{bmatrix}^{-1} \\ &= \begin{bmatrix}\left(\hat{\sigma}^{-2}\sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top(X_g - \tilde{X}_g)\right)^{-1} & 0 \\ 0 & \left(\sum_{g=1}^{G}\tilde{X}_g^\top V_g^{-1}\tilde{X}_g\right)^{-1}\end{bmatrix} \qquad (78)\end{aligned}$$

Because of Equation (72), this means, like in the FE setting, that the only portion of $M^{(\mathrm{bcMLM})}$ that contributes to the estimated variance of $\hat{\beta}$ in the bcMLM setting is its first $p-1$ rows and columns, i.e.,

$$\begin{aligned}&(X - \tilde{X})^\top V^{-1}\begin{bmatrix}\hat{e}_1^{(\mathrm{bcMLM})}[\hat{e}_1^{(\mathrm{bcMLM})}]^\top & & 0 \\ & \ddots & \\ 0 & & \hat{e}_G^{(\mathrm{bcMLM})}[\hat{e}_G^{(\mathrm{bcMLM})}]^\top\end{bmatrix}V^{-1}(X - \tilde{X}) \\ &\qquad = \sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top \hat{V}_g^{-1}\hat{e}_g^{(\mathrm{bcMLM})}[\hat{e}_g^{(\mathrm{bcMLM})}]^\top \hat{V}_g^{-1}(X_g - \tilde{X}_g) \\ &\qquad = \sum_{g=1}^{G}\left((X_g - \tilde{X}_g)^\top \hat{V}_g^{-1}\hat{e}_g^{(\mathrm{bcMLM})}\right)\left((X_g - \tilde{X}_g)^\top \hat{V}_g^{-1}\hat{e}_g^{(\mathrm{bcMLM})}\right)^\top \qquad (79)\end{aligned}$$

Expanding $(X_g - \tilde{X}_g)^\top \hat{V}_g^{-1}\hat{e}_g^{(\mathrm{bcMLM})}$ in Equation (79) using the definition of $\hat{e}_g^{(\mathrm{bcMLM})}$ and the expression for V_g^{-1} in Equation (75), we get

$$\begin{aligned}(X_g - \tilde{X}_g)^\top \hat{V}_g^{-1}\hat{e}_g^{(\mathrm{bcMLM})} &= (X_g - \tilde{X}_g)^\top(\hat{\sigma}^{-2}I_{n_g} - Z_g\hat{A}_g Z_g^\top)[Y_g - (X_g - \tilde{X}_g)\hat{\beta} - \tilde{X}_g\hat{\alpha}] \\ &= \hat{\sigma}^{-2}(X_g - \tilde{X}_g)^\top[Y_g - (X_g - \tilde{X}_g)\hat{\beta} - \tilde{X}_g\hat{\alpha}] \\ &= \hat{\sigma}^{-2}(X_g - \tilde{X}_g)^\top(Y_g - X_g\hat{\beta}) \qquad (80)\end{aligned}$$

Plugging Equation (80) above into Equation (79), the first $p-1$ rows and columns of $M^{(\mathrm{bcMLM})}$, yields

$$\hat{\sigma}^{-4}\sum_{g=1}^{G}(X_g - \tilde{X}_g)^\top(Y_g - X_g\hat{\beta})(Y_g - X_g\hat{\beta})^\top(X_g - \tilde{X}_g) \qquad (81)$$

Therefore, the CRSEs for $\hat{\beta}$ from bcMLM are

$$\begin{aligned}\widehat{\mathrm{var}}^{(\mathrm{bcMLM})}_{\mathrm{CRSE}}(\hat{\beta}) = c \times & \Big(\hat{\sigma}^{-2} \sum_{g=1}^{G} (X_g - \tilde{X}_g)^{\top} (X_g - \tilde{X}_g)\Big)^{-1} \\ & \times \Big(\hat{\sigma}^{-4} \sum_{g=1}^{G} (X_g - \tilde{X}_g)^{\top} (Y_g - X_g \hat{\beta})(Y_g - X_g \hat{\beta})^{\top} (X_g - \tilde{X}_g)\Big) \\ & \times \Big(\hat{\sigma}^{-2} \sum_{g=1}^{G} (X_g - \tilde{X}_g)^{\top} (X_g - \tilde{X}_g)\Big)^{-1} \\ = c \times & \Big(\sum_{g=1}^{G} (X_g - \tilde{X}_g)^{\top} (X_g - \tilde{X}_g)\Big)^{-1} \\ & \times \Big(\sum_{g=1}^{G} (X_g - \tilde{X}_g)^{\top} (Y_g - X_g \hat{\beta})(Y_g - X_g \hat{\beta})^{\top} (X_g - \tilde{X}_g)\Big) \\ & \times \Big(\sum_{g=1}^{G} (X_g - \tilde{X}_g)^{\top} (X_g - \tilde{X}_g)\Big)^{-1} \end{aligned} \tag{82}$$

which, if c is chosen to be the same in bcMLM and FE as we recommend in Appendix A.15, is exactly equal to $\widehat{\mathrm{var}}^{(\mathrm{FE})}_{\mathrm{CRSE}}(\hat{\beta})$ in Equation (71). Therefore, the CRSEs from both FE and bcMLM for $\hat{\beta}$ are equivalent.

□