

Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Time-Varying Treatments

Stephen W. Raudenbush
The University of Chicago

April 7, 2008

Prepared for the National Conference on Value-Added Modeling
April 22-24, 2008
University of Wisconsin at Madison

The work reported here was supported by funds from the Spencer Foundation for the project “Improving Research on Instruction: Models, Designs, and Analytic Methods.”

Please do not cite or quote this draft without the permission of the author.

Abstract

Of widespread interest in social science are observational studies in which entities (persons, schools, states, countries, etc.) are exposed to varied treatment conditions over time. As in all observational studies, the non-randomized assignment of treatments poses challenges to valid causal inference. An attractive feature of panel studies with time-varying treatments, however, is that the design makes it possible to remove the influence of unobserved time-invariant confounders in assessing the impact of treatments. The removal of such confounding is typically achieved by including fixed effects of the repeatedly measured entities. In some cases, these entities are clustered in larger units, and the time-invariant influences of these larger units can be removed by adding additional dimensions of fixed effects; also, entity-invariant temporal influences can be removed through time fixed effects. In this paper, I introduce an alternative procedure: *adaptive centering of treatment variables with random effects*. I show that this alternative procedure can replicate the fixed effects analysis of time-varying treatments in any dimension of clustering and offers several comparative advantages: the incorporation into standard errors of multiple sources of uncertainty; the modeling of heterogeneity of treatment effects; estimation of treatment effects at multiple levels and their interaction; improved estimates of unit-specific effects; and computational simplicity. The paper shows how adaptive centering can efficiently estimate effects in an L -level model using one or more dimensions of clustering.

1. Introduction

Consider an L -level linear statistical model with random intercepts defined at every level and a treatment variable Z that varies at level $L-1$. Adaptive centering involves centering Z and any covariates X around their group-level means defined at level L . The mean around which Z is centered is a weighted average using as a weight the precision of the outcome sample mean for each level $L-1$ unit. Estimation of this L -level random intercept model will remove the influence of all L -level confounding variables, observed and unobserved. The random effects structure will account for clustering at levels $1, \dots, L-1$. An example is a three-level model with students (level-1) nested within classrooms (level 2) and classrooms nested within schools (level cross -3), where the treatment is at the classroom level. Each classroom-level treatment indicator is centered around its school mean; each school mean is a weighted average, with weights equal to the precision with which each classroom's outcome mean is estimated. This approach will remove all confounding by unobserved school-level covariates while efficiently using information about variation within and between classrooms to obtain efficient point estimates and standard errors that account for clustering within schools.

The two-level model is a special case, in which the treatment Z varies at level-1. An example involves repeated measures at level 1 nested within persons at level 2. The persons are exposed to varied values of a treatment variable over time, and the aim is to remove all person-specific time-invariant confounding from the error with which the treatment effects are estimated. Under standard assumptions, the adaptive centering approach will center the treatment variable Z around its person-specific mean. This will replicate the standard econometric analysis that uses student fixed effects with ordinary least squares (OLS) regression (Green, 1997).

To incorporate time-varying treatments with clustered data, consider now a setting in which students are flowing across teachers and the treatment is a time-varying Z . This involves a crossed-nested structure in which time-series observations are nested within a cross-classification of students and teachers (Raudenbush and Bryk, Chapter 12, Example 2). The aim is to remove unobserved confounding defined on students and teachers. This requires centering the treatment indicator around both student and the teacher means. The adaptive centering approach replicates the econometric approach that uses student and teacher fixed effects with OLS regression (Green, 1997). Now suppose that the teachers of interest are nested within schools, and the treatment Z is a teacher characteristic. Thus, the observations are crossed by students and teachers and the teachers are nested within schools. The aim is to remove time-invariant confounding associated with students and schools and also to insure that that the standard error of the treatment effect incorporates the clustering of observations within the student-by-teacher crossed classification. Unlike the standard fixed effects approach, the adaptive centering provides a natural analysis. The treatment variable will now be centered around "predicted values" based on student and school main effects; however, these main effects will now be weighted estimates using a precision-weighting scheme that reflect the clustering by student and teacher.

The approach can be extended to an arbitrary number of levels and dimensions of centering. The estimates based on the L -level model with adaptive centering and maximum likelihood estimation will possess the same asymptotic efficiency as an $L-1$ level hierarchical linear model estimated via maximum likelihood.

This chapter begins by considering the causal effects that can be estimated using the fixed effects approach in the context of a time-varying treatment. Section 3 illustrates how the adaptive centering approach works based on a simple hypothetical example. Section 4 provides a general L -level model, clarifies the assumptions that must be met for valid inference and provides a general set of equations to define adaptive centering. Section 5 takes up the case of one-dimensional centering, first for the general L -level model, then with application to two-level and three-level models. Section 6 considers cross-classified models with two dimensions of centering, focusing on the computational challenge.

2. Time-Varying Treatment Effects Under a Fixed Effects Model

Perhaps the most common application of fixed effects methods for causal analysis involves time-varying treatments within an entity, most often a person. A typical specification for the observed-data model is

$$y_{it} = \beta x_{it} + u_i + \varepsilon_{it} \quad (2.1)$$

where y_{it} is a continuous outcome measured for person i at time t , $t=0,1,\dots,T_i$; $i=1,\dots,n$; x_{it} is a measure of treatment received; u_i is an unobserved person-specific fixed effect capturing all time-invariant characteristics of person i that predict the outcome; and ε_{it} is a random error uncorrelated with u_i and x_{it} .

What causal effect might β represent in (1)? We can answer this question for a binary treatment and $T=2$ and then for any T by induction. Suppose that no treatments are implemented at $t=0$, but that treatment z_1 is implemented at $t=1$ and treatment z_2 is implemented at $t=2$. For simplicity we shall regard these as binary treatments, so that person i might receive the sequence (z_1, z_2) with $(0,0)$ representing no exposure to the treatment, $(1,0)$ represents exposure at time 1 but not time 2; $(0,1)$ representing exposure at time 2 but not time 1; and $(1,1)$ representing exposure at times 1 and 2. Under the “stable unit-treatment value assumption” or “SUTVA” (Rubin, 1986), each person has two potential outcomes, $y_1(z_1)$, $z_1 \in \{0,1\}$ at time 1 and four potential outcomes $y_2(z_1, z_2)$, $z_1 \in \{0,1\}$, $z_2 \in \{0,1\}$ at time 2. Define the average causal effect of receiving $z_1=1$ at $t=1$ as $E[Y_1(1) - Y_1(0)] = \delta_{11}$ and define three causal effects at time 2:

$$\begin{aligned} E[Y_2(1,0) - Y_2(0,0)] &= \delta_{21} \\ E[Y_2(0,1) - Y_2(0,0)] &= \delta_{22} \\ E[Y_2(1,1) - Y_2(0,0)] &= \delta_{21} + \delta_{22} + \delta^* \end{aligned} \quad (2.2)$$

so that δ_{11} is the average effect of time 1 treatment on time 1 outcome; δ_{21} is the average effect of time 1 treatment on time 2 outcome if no treatment is received at time 2; δ_{22} is the average effect of time 2 treatment on time 2 outcome if no treatment is received at time 2; and δ^* an “amplifying” effect of receiving the treatment at both times, equivalent to the statistical interaction effect of z_1 and z_2 (Hong and Raudenbush, in press). We can summarize the causal effects (2) as

$$\begin{aligned} E[Y_1(z_1)] &= Y(0) + z_1\delta_{11} \\ E[Y_2(z_1, z_2)] &= Y(0,0) + z_1\delta_{21} + z_2\delta_{22} + z_1z_2\delta^*. \end{aligned} \quad (2.3)$$

For $T=2$ and binary treatments, these four causal effects “saturate” the causal effect space, so that any causal effect estimated by (1) must be a subset of these four. Two sensible simplifications come to mind, the “cumulative effects model” and the “ephemeral effects model.”

2.1 Cumulative effects

Under the cumulative effects model,

$$\delta_{11} = \delta_{21} = \delta_{22} = \beta; \delta^* = 0 \quad (2.4)$$

We therefore find

$$E[Y_1(z_1)] = E[Y_1(0)] + \beta z_1 \quad (2.5)$$

And

$$E[Y(z_1, z_2)] = E[Y(0,0)] + \beta(z_1 + z_2). \quad (2.6)$$

Assume further the linear structure

$$\begin{aligned} Y_{1i}(z_{1i}) &= \beta z_{1i} + u_i + e_{1i}(z_{1i}) \\ Y_{2i}(z_{1i}, z_{2i}) &= \beta(z_{1i} + z_{2i}) + u_i + e_{2i}(z_{1i}, z_{2i}), \end{aligned} \quad (2.7)$$

a special form of (1) with

$$x_{1i} = z_{1i}, \quad x_{2i} = z_{2i} + z_{1i} \quad (2.8)$$

Now let us envision a study in which the assigned treatment at $t=1$ is Z_1 , a random variable taking on value of $Z_1 = z_1$ and the assigned treatment at $t=2$ is Z_2 taking on the value $Z_2 = z_2$. The observed outcomes are therefore $Y_1 = Z_1Y_1(1) + (1 - Z_1)Y_1(0)$ at time 1 and $Y_2 = Z_1Z_2Y_2(1,1) + (1 - Z_1)Z_2Y_2(0,1) + Z_1(1 - Z_2)Y_2(1,0) + (1 - Z_1)(1 - Z_2)Y_2(0,0)$ at time 2.

We now assume a special form of the “strongly ignorable treatment assignment,” (Rosenbaum and Rubin, 1983). In this case, we assume only that treatment assignment is independent of the within-subject random effects. That is

$$Z_1, Z_2 \perp e_1 = Z_1 e_1(1) + (1 - Z_1) e_1(0) \quad (2.9)$$

and

$$\begin{aligned} Z_1, Z_2 \perp e_2 = & Z_1 Z_2 e_2(1,1) + (1 - Z_1) Z_2 e_2(0,1) \\ & + Z_1 (1 - Z_2) e_2(1,0) + (1 - Z_1) (1 - Z_2) e_2(0,0). \end{aligned} \quad (2.10)$$

The key benefit of the fixed effects specification is that we do not need to make an assumption about the ignorability of person-specific effects u . This results from a benefit of the fixed effects specification. Estimation of β in (1) will involve within-person deviations in y , removing u from the estimation of causal effects. Specifically, the point estimate will be

$$\hat{\beta} = \frac{\sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i) (y_{it} - \bar{y}_i)}{\sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)^2}, \quad (2.11)$$

having expectation

$$\begin{aligned} E(\hat{\beta} | Z_{1i} = z_{1i}, Z_{2i} = z_{2i}) \\ = \beta + \frac{\sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i) E(e_{it} - \bar{e}_i | Z_{1i} = z_{1i}, Z_{2i} = z_{2i})}{\sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)^2}. \end{aligned} \quad (2.12)$$

Under our ignorability assumption (11), the second term in the numerator becomes null and the estimator is unbiased.

A key assumption, of course, is that removal of time-invariant confounding captured in u is sufficient to remove bias. This assumption can be relaxed by adding observed time-varying covariates as long as those time-varying covariates are not influenced by prior treatments. Robins (2000) developed inverse probability of treatment weighting (IPTW) as a strategy for adjusting for such time-varying confounders. Hong and Raudenbush (in press) extended this to time-varying treatments where students and teachers are nested within schools. Robbins’ approach assumes sequentially strongly ignorable treatment assignment. In one way, assumption (2.10) is stronger than required under sequential ignorability. Assumption (2.10) requires that the potential random errors at $t=1$ be independent not only of treatment assignment at $t=1$ but also of treatment assignment at $t=2$.

2.2 Ephemeral Effects

Under the “ephemeral effects” model, treatment effects decay over time so that

$$\delta_{12} = \delta_{22} = \beta; \delta_{21} = \delta^* = 0. \quad (2.13)$$

We therefore find

$$E[Y_1(z_1)] = E[Y_1(0)] + \beta z_1 \text{ and} \quad (2.14)$$

And

$$E[Y(z_1, z_2)] = E[Y(0,0)] + \beta z_2. \quad (2.15)$$

By again adopting the assumptions of a linear structure and strongly ignorable treatment assignment given person fixed effects u , (2.1) identifies the causal effects (2.13) with

$$x_{1i} = z_{1i}, \quad x_{2i} = z_{2i} \quad (2.16)$$

2.3 Generalizations

Other models for causal effects. The cumulative and ephemeral effects models represent the two extremes on a continuum. There is no decay in the causal effect at any time in the cumulative effects model while there is total decay in the ephemeral effects model. An intermediate model would parameterize the rate of decay to lie between these two extremes. See McCaffrey et al (2003) in the context of value added models.

Multi-dimensional fixed effects. Model (2.1) is readily extended to multiple dimensions of fixed effects. For example, students may move across teachers over time, and causal effects may be embedded in a model that specifies fixed effects of students and fixed effects of teachers or schools, as illustrated in the next section.

Non-binary treatments. The logic described here can readily be extended to multi-valued treatments such as multinomial treatments and continuous dosages.

3. Hypothetical Example

To illustrate the adaptive centering approach, consider the hypothetical data set in Table 1. We have 20 “children” $i=1, \dots, n=20$ (rows of the table) each observed on three occasions, with one occasion in each of three “schools” $k=1, \dots, K=3$ (see the three major columns). Nested within each school are three “teachers” $j=1, \dots, J_k=3$, so that there are nine teachers over all. The treatment variable $x \in \{-1, 0, 1\}$ is a school characteristic, though it varies within children as they move across schools.

Insert Table 1 About Here

The data were generated according to the model

$$y_{ijk} = \theta + \beta x_{tik} + u_i + s_k + \varepsilon_{ijk} \quad (3.1)$$

where

$$\begin{aligned} u_i &= \gamma w_i + \phi(\text{childid})_i \\ s_k &= \delta(\text{schoolid} - 2)_k. \end{aligned}$$

with

$$\theta = 0; \quad \beta = 2; \quad \gamma = 5; \quad \delta = 4; \quad \phi = .5; \quad \varepsilon_{ijk} \sim N(0,1).$$

The errors ε are mutually independent and independent of the other elements of the model.

In this scenario, w_i is unobserved and the researcher is unaware of the fact that linear functions of the *childid* and *schoolid* contribute to the outcome. The central aim is to estimate $\beta = 2$ using the observed $y, x, \text{childid}, \text{schoolid}$. The fact that child and school effects are correlated with treatment x invalidates the assumption of the standard random effects model when u_i, s_k are regarded as random, that is

$$\begin{aligned} E(y_{ijk} | X = x_{tik}) &= \theta + \beta x_{ti} + E(u_i + s_k | X = x_{tik}) \\ &= \theta + \beta x_{tik} + E(u_i + s_k) \\ &= \theta + \beta x_{tik} \end{aligned} \quad (3.2)$$

The failure of assumption (3.2) implies that estimation of the random effects model will produce a biased estimate of β .

3.1 One-Dimensional Confounding

One-dimensional fixed effects model. Suppose first that the analyst wishes to control for time-invariant child differences but ignores the possibility of time-invariant school-level confounding. Therefore, this analyst fits model (1) yielding the estimates in Table 2. The estimate $\hat{\beta} = 5.498$, $se = 0.856$ is far off the mark of $\beta = 2$, reflecting the failure to control for school-level confounding.

Insert Table 2 About Here

Adaptive centering with random effects. As an alternative, considering the random effects model

$$y_{ijk} = \theta + \beta(x_{tik} - \bar{x}_i) + u_i + \varepsilon_{ijk} \quad (3.3)$$

where

$$u_i \sim N(0, \tau^2), \quad \varepsilon_{ijk} \sim N(0, \sigma^2),$$

$$\bar{x}_{\cdot i} = \sum_{t=1}^3 x_{tik} / 3.$$

I estimated (3.3) by maximum likelihood using the HLM6 program (Raudenbush, Bryk, Cheong, and Congdon, 2006). Inferences (Table 3) regarding β are identical to those based on the fixed effects model (2.1) with no centering of x . A question of interest in the next section will involve when and why these equivalences will hold.

Insert Table 3 About Here

3.2 Two-Dimensional Confounding

Two-dimensional fixed effects model. Suppose now that the analyst decides to control for time-invariant school level differences as well as for time-invariant child differences using the two dimensional fixed effects model. The estimates are given in Table 4. We see that $\hat{\beta} = 2.57$, $se = 0.288$ is now within the vicinity of the true $\beta = 2$, reflecting the benefit of controlling for school-level confounding in addition to student-level confounding.

Insert Table 4 About Here

Adaptive centering with random effects. Now consider the alternative random effects model with two dimensional centering

$$y_{ijk} = \theta + \beta(x_{tik} - \bar{x}_{\cdot i} - \bar{x}_{\cdot k} + \bar{x}_{\cdot\cdot}) + u_i + s_k + \varepsilon_{ijk} \quad (3.4)$$

where

$$u_i \sim N(0, \tau^2), \quad s_k \sim N(0, \omega^2), \quad \varepsilon_{ijk} \sim N(0, \sigma^2),$$

$$\bar{x}_{\cdot i} = \sum_{t=1}^3 x_{tik} / 3, \quad \bar{x}_{\cdot k} = \sum_{i=1}^{20} x_{tik} / 20$$

I estimated (3.4) by maximum likelihood. Inferences (Table 5) regarding β are identical to those based on the two-dimensional fixed effects model (15) with no centering of x .

Insert Table 5 About Here

3.3 A Richer Class of Models

The results of this hypothetical example suggest that adaptive centering of treatment indicators with random effects can replicate the fixed effects estimates in any dimension. In fact, within the random effects framework, a richer class of models can be estimated.

Accounting for uncertainty. For example, it is possible to estimate a random effect for each teacher in the context of our example:

$$y_{ijk} = \theta + \beta(x_{tik} - \bar{x}_{\cdot i} - \bar{x}_{\cdot k} + \bar{x}_{\cdot\cdot}) + u_i + c_{j(k)} + s_k + \varepsilon_{ijk} \quad (3.5)$$

where we add the additional classroom random effect $c_{j(k)} \sim N(0, \psi^2)$. Specification of this random effect allows the analysis to incorporate uncertainty associated with classrooms, presumably providing more realistic standard errors than when such clustering is ignored. An alternative method for obtaining consistent standard errors is the Huber-White approach (Huber and White, 1967). However, that approach would require multiplying the vector of residuals of every student with the vectors of residuals of every other student, given that the mobility of students over schools and teachers induces covariances of residuals across all students. This would be computationally difficult in large applications.

Heterogeneous treatment effects. Moreover, it is straightforward within the random effects framework to allow random coefficients. Consider the model

$$\begin{aligned} y_{ijk} &= \theta + u_{0i} + s_{0k} + c_{0j(k)} + (\beta + u_{1ik} + s_{1k})(x_{tik} - \bar{x}_i - \bar{x}_k + \bar{x}) + \varepsilon_{ijk} \\ \begin{bmatrix} u_{0i} \\ u_{1ik} \end{bmatrix} &\sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right] \\ \begin{bmatrix} s_{0k} \\ s_{1k} \end{bmatrix} &\sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{pmatrix} \right] \\ c_{j(k)} &\sim N(0, \psi^2) \\ \varepsilon_{ijk} &\sim N(0, \sigma^2). \end{aligned} \quad (3.6)$$

The variance components τ_{11}, ω_{11} parameterize the heterogeneity of the treatment effect across children and schools.

Multilevel factorial designs. We can also readily extend the random effects approach to incorporate multi-level treatments and cross-level interactions. Consider the case of a between-school characteristic or treatment that interacts with x :

$$\begin{aligned}
y_{ijk} &= \theta + u_{0i} + s_{0k} + c_{0j(k)} + (\beta + u_{1ik} + s_{1k})(x_{iik} - \bar{x}_i - \bar{x}_k + \bar{x}) + \\
&\quad + \gamma_0 w_k + \gamma_1 w_k^* (x_{iik} - \bar{x}_i - \bar{x}_k + \bar{x}) + \varepsilon_{ijk} \\
\begin{bmatrix} u_{0i} \\ u_{1ik} \end{bmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right] \\
\begin{bmatrix} s_{0k} \\ s_{1k} \end{bmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{pmatrix}\right] \\
c_{j(k)} &\sim N(0, \psi^2) \\
\varepsilon_{ijk} &\sim N(0, \sigma^2)
\end{aligned} \tag{3.7}$$

In this case γ_0 is the main effect of the between-school predictor w and γ_1 is the cross-level interaction effect. Such specifications are not possible within the fixed effects framework.

4. General Model

We begin with the general linear mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{A}\mathbf{b} + \mathbf{e}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Omega}), \quad \mathbf{e} \sim N(\mathbf{0}, \mathbf{V}^*) \tag{4.1}$$

where \mathbf{Y} is vector of outcomes; $\boldsymbol{\gamma} = (\theta, \boldsymbol{\beta})^T$ is a vector of unknown fixed regression coefficients; \mathbf{b} and \mathbf{e} are vectors of unknown random effects, $\mathbf{X} = [1 \quad \mathbf{x}]$ and \mathbf{A} are known design matrices dimensioned conformably; and $\boldsymbol{\Omega}$ and \mathbf{V}^* are positive definite covariance matrices. In the case of $\boldsymbol{\Omega}$ and \mathbf{V}^* known, the maximum likelihood estimator of the regression coefficients is

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y} \\
&= \begin{bmatrix} \hat{\theta} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \mathbf{V}^{-1} \mathbf{1} & \mathbf{1}^T \mathbf{V}^{-1} \mathbf{x} \\ \mathbf{x}^T \mathbf{V}^{-1} \mathbf{1} & \mathbf{x}^T \mathbf{V}^{-1} \mathbf{x} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}^T \mathbf{V}^{-1} \mathbf{Y} \\ \mathbf{x}^T \mathbf{V}^{-1} \mathbf{Y} \end{bmatrix}
\end{aligned} \tag{4.2}$$

where $\mathbf{V} = \text{Var}(\mathbf{Y}) = \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T + \mathbf{V}^*$ and (Lindley and Smith, 1972)

$$\mathbf{V}^{-1} = (\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T + \mathbf{V}^*)^{-1} = \mathbf{V}^{*-1} - \mathbf{V}^{*-1} \mathbf{A} (\mathbf{A}^T \mathbf{V}^{*-1} \mathbf{A} + \mathbf{V}^{*-1})^{-1} \mathbf{A}^T \mathbf{V}^{*-1}.$$

A key assumption is that of no association between the random effects \mathbf{b} , \mathbf{e} and the predictors, \mathbf{x} , in which case

$$\begin{aligned}
E(\hat{\boldsymbol{\gamma}} | \mathbf{x}) &= (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} E[\mathbf{Y} | \mathbf{x}] \\
&= \boldsymbol{\gamma} + (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} [\mathbf{A}E(\mathbf{b} | \mathbf{x}) + E(\mathbf{e} | \mathbf{x})] \\
&= \boldsymbol{\gamma} + (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} [\mathbf{A}E(\mathbf{b}) + E(\mathbf{e})] \\
&= \boldsymbol{\gamma}.
\end{aligned} \tag{4.3}$$

We are willing to stipulate independence of \mathbf{e} and \mathbf{x} so that $E(\mathbf{e} | \mathbf{x}) = E(\mathbf{e}) = \mathbf{0}$ but not the independence of \mathbf{b} and \mathbf{x} . So (4.3) is not generally applicable.

The task of centering is then to solve the equations

$$\begin{aligned}
\mathbf{1}^T \mathbf{V}^{*-1} \mathbf{x} &= \mathbf{0} \\
\mathbf{x}^T \mathbf{V}^{*-1} \mathbf{A} &= \mathbf{0}
\end{aligned} \tag{4.4}$$

in which case

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^T \mathbf{V}^{*-1} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{V}^{*-1} \mathbf{Y} \tag{4.5}$$

from which it follows that

$$E(\hat{\boldsymbol{\beta}} | \mathbf{x}) = \boldsymbol{\beta} + (\mathbf{x}^T \mathbf{V}^{*-1} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{V}^{*-1} \mathbf{A}E(\mathbf{b} | \mathbf{x}) = \boldsymbol{\beta} \tag{4.6}$$

even if $E(\mathbf{b} | \mathbf{x}) \neq \mathbf{0}$. We show in the next section how to solve these equations in the case of one-dimensional confounding (e.g., confounding of time-invariant child effects) and two dimensional confounding (e.g., time-invariant child and school effects). The approach extends, of course, to multiple dimensions of confounding. We consider an arbitrary number of levels of nesting and crossing within treatments.

5. Specific Sub-Models

The fixed effects specification may be regarded as enabling the researcher to simulate a randomized block design. Between-block confounding is removed through specification of fixed block effects as explained in Section 3. Within blocks, treatments and potential outcomes are regarded as independent, possibly given specification of certain within-block covariates. I use the term ‘‘one-dimensional confounding’’ to describe the case in which the blocks are arrayed on a single dimension. In the illustrative example of Section 3, the blocks were children (a single dimension). In Section 4, I illustrated a case in which adaptive centering of the within-child explanatory variables with random child effects replicated the fixed effects specification. However, that section also considered the problem of two-dimensional confounding, where blocking was based on children and schools. That section illustrated a case in which adaptively centering the treatment variable around child and school means replicated a two-dimensional fixed effects specification.

In this Section, I show how to derive the appropriate adaptive centering for one- and two-dimensional confounding. The basic idea is that there are one or two dimensions of blocking; within blocks, potential outcomes and explanatory variables are conditionally independent given observed within-block covariates. Within blocks, there may be an arbitrary number of layers of nested and crossed random factors.

5.1 One Dimensional Confounding

We now consider the case in which there is an L -level nested structure with a treatment variable defined at level $L-1$. Units at level L are regarded as blocks and the aim is to remove block effects from the experimental error by means of adaptive centering. The question is how to weight the observations so that the block effects are removed. After showing the general result for the L -level model, I shall illustrate how the procedure works for two-level models ($L=2$) and three-level models ($L=3$).

L-level model. Our model is

$$\mathbf{Y}_r = \mathbf{X}_r \boldsymbol{\gamma} + \mathbf{1}_r \mathbf{b}_r + \mathbf{e}_r, \quad \mathbf{b}_r \sim N(\mathbf{0}, \omega^2 \mathbf{I}_r), \quad \mathbf{e}_r \sim N(\mathbf{0}, \mathbf{V}_{L-1,r}). \quad (5.1)$$

Here \mathbf{Y}_r is the n_r by 1 vector of outcomes within L -level unit r having elements $\{y_{ijk\dots r}\}$. The regression coefficient vector $\boldsymbol{\gamma} = (\theta, \boldsymbol{\beta})^T$ includes an intercept θ and a regression coefficient vector $\boldsymbol{\beta}$. Correspondingly, $\mathbf{X}_r = [\mathbf{1} \quad \mathbf{x}_r]$ is the known design matrix for the fixed effects; \mathbf{I}_r is the n_r by n_r identity matrix. In the case of one-dimensional blocking, the random effects design is simply the n_r by 1 vector $\mathbf{1}_r$, all of whose elements are unity. The covariance matrix $\mathbf{V}_{L-1,r} = \text{Var}(\mathbf{Y}_r | \mathbf{b}_r)$ and $\mathbf{V}_{L,r} = \text{Var}(\mathbf{Y}_r) = \omega^2 \mathbf{1}_r \mathbf{1}_r^T + \mathbf{V}_{L-1,r}$.

In the case of $\boldsymbol{\Omega}$ and $\mathbf{V}_{L-1,r}$ known, the maximum likelihood estimator of the regression coefficients is

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \left(\sum_{r=1}^R \mathbf{X}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{X}_r \right)^{-1} \sum_{r=1}^R \mathbf{X}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{Y}_r \\ &= \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \sum_{r=1}^R \mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{1}_r & \sum_{r=1}^R \mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{x}_r \\ \sum_{r=1}^R \mathbf{x}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{1}_r & \sum_{r=1}^R \mathbf{x}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{x}_r \end{bmatrix}^{-1} \begin{bmatrix} \sum_{r=1}^R \mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{Y}_r \\ \sum_{r=1}^R \mathbf{x}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{Y}_r \end{bmatrix} \end{aligned} \quad (5.2)$$

We now set

$$x_r^* = x_r - \mathbf{1}_r \bar{x}_r \quad (5.3)$$

where

$$\bar{x}_r = (\mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{1}_r)^{-1} \mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{x}_r. \quad (5.4)$$

We then have

$$\mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{x}_r^* = \mathbf{0}. \quad (5.5)$$

Substituting \mathbf{x}_r^* for \mathbf{x}_r in model (5.1) and estimating equation (5.2), we now find that

$$\begin{aligned} \sum_{r=1}^R \mathbf{1}_r^T \mathbf{V}_{L,r}^{-1} \mathbf{x}_r^* &= \sum_{r=1}^R \mathbf{1}_r^T (\omega^2 \mathbf{1}_r \mathbf{1}_r^T + \mathbf{V}_{L-1,r})^{-1} \mathbf{x}_r^* \\ &= \sum_{r=1}^R \mathbf{1}_r^T \left[\mathbf{V}_{L-1,r}^{-1} - \mathbf{V}_{L-1,r}^{-1} \mathbf{1}_r (\mathbf{1}_r^T \mathbf{V}_{L-1,r} \mathbf{1}_r + \omega^{-2} \mathbf{I}_r)^{-1} \mathbf{1}_r^T \right] \mathbf{x}_r^* \\ &= \sum_{r=1}^R \mathbf{1}_r^T \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* - \sum_{r=1}^R \mathbf{1}_r^T \mathbf{V}_{L-1,r}^{-1} \mathbf{1}_r (\mathbf{1}_r^T \mathbf{V}_{L-1,r} \mathbf{1}_r + \omega^{-2} \mathbf{I}_r)^{-1} \mathbf{1}_r^T \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* \\ &= \mathbf{0} \end{aligned} \quad (5.6)$$

$$\begin{aligned} \sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L,r}^{-1} \mathbf{x}_r^* &= \sum_{r=1}^R \mathbf{x}_r^{*T} (\omega^2 \mathbf{1}_r \mathbf{1}_r^T + \mathbf{V}_{L-1,r})^{-1} \mathbf{x}_r^* \\ &= \sum_{r=1}^R \mathbf{x}_r^{*T} \left[\mathbf{V}_{L-1,r}^{-1} - \mathbf{V}_{L-1,r}^{-1} \mathbf{1}_r (\mathbf{1}_r^T \mathbf{V}_{L-1,r} \mathbf{1}_r + \omega^{-2} \mathbf{I}_r)^{-1} \mathbf{1}_r^T \right] \mathbf{x}_r^* \\ &= \sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* - \sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{1}_r (\mathbf{1}_r^T \mathbf{V}_{L-1,r} \mathbf{1}_r + \omega^{-2} \mathbf{I}_r)^{-1} \mathbf{1}_r^T \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* \\ &= \sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* \end{aligned} \quad (5.7)$$

Using a similar argument,

$$\sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L,r}^{-1} \mathbf{Y}_r = \sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{Y}_r \quad (5.8)$$

With these results in mind, we can see that, using the centered value of $x_r^* = x_r - 1_r \bar{x}_r$ as a predictor,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* \right)^{-1} \sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{Y}_r \\ \text{Var}(\hat{\boldsymbol{\beta}}) &= \left(\sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{V}_{L-1,r}^{-1} \mathbf{x}_r^* \right)^{-1}. \end{aligned} \quad (5.9)$$

Two-level model. The two-level model is a special case of (5.1) with $L=2$, $r=j$, so that \mathbf{Y}_j is the n_j by 1 vector of outcomes within level-2 unit j having elements $\{y_{ij}\}$. The covariance matrix $\mathbf{V}_{1,j} = \text{Var}(\mathbf{Y}_j | \mathbf{b}_j) = \sigma^2 \mathbf{I}_j$ and $\mathbf{V}_{2,j} = \text{Var}(\mathbf{Y}_j) = \omega^2 \mathbf{1}_j \mathbf{1}_j^T + \sigma^2 \mathbf{I}_j$ with σ^2 denoting the level-1 variance. For example, j may denote the person, i the time point; σ^2 is the within-person variance, and ω^2 is the between-person variance

In the case of ω^2 and σ^2 known, the maximum likelihood estimator of the regression coefficients is

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \left(\sum_{j=1}^J \mathbf{X}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{X}_j \right)^{-1} \sum_{j=1}^J \mathbf{X}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{Y}_j \\ &= \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^J \mathbf{1}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{1}_j & \sum_{j=1}^J \mathbf{1}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{x}_j \\ \sum_{j=1}^J \mathbf{x}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{1}_j & \sum_{j=1}^J \mathbf{x}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{x}_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^J \mathbf{x}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{Y}_j \\ \sum_{j=1}^J \mathbf{x}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{Y}_j \end{bmatrix} \end{aligned} \quad (5.10)$$

We now set

$$\mathbf{x}_j^* = \mathbf{x}_j - \mathbf{1}_j \bar{x}_j \quad (5.11)$$

where

$$\bar{x}_j = (\mathbf{1}_j^T \mathbf{1}_j)^{-1} \mathbf{1}_j^T \mathbf{x}_j = \sum_{i=1}^{n_j} x_{ij} / n_j. \quad (5.12)$$

We then have

$$\mathbf{1}_j^T \mathbf{x}_j^* = \mathbf{0}. \quad (5.13)$$

Substituting \mathbf{x}_j^* for \mathbf{x}_j in (5.10), we now find that

$$\begin{aligned} \sum_{j=1}^J \mathbf{1}_j^T \mathbf{V}_{2,j}^{-1} \mathbf{x}_j^* &= \sum_{j=1}^J \mathbf{1}_j^T (\omega^2 \mathbf{1}_j \mathbf{1}_j^T + \sigma^2 \mathbf{I}_j)^{-1} \mathbf{x}_j^* \\ &= \sum_{j=1}^J \sigma^{-2} \mathbf{1}_j^T \left[\mathbf{I}_j - \mathbf{1}_j (\mathbf{1}_j^T \mathbf{1}_j + \sigma^2 \omega^{-2}) \mathbf{1}_j^T \right] \mathbf{x}_j^* \\ &= \sigma^{-2} \sum_{j=1}^J \mathbf{1}_j^T \mathbf{x}_j^* - \sum_{r=1}^R \mathbf{1}_j^T \mathbf{1}_j (\mathbf{1}_j^T \mathbf{1}_j + \sigma^2 \omega^{-2}) \mathbf{1}_j^T \mathbf{x}_j^* \\ &= \mathbf{0} \end{aligned} \quad (5.14)$$

$$\begin{aligned}
\sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{V}_{2,j}^{-1} \mathbf{x}_j^* &= \sum_{j=1}^J \mathbf{x}_j^{*T} (\omega^2 \mathbf{1}_j \mathbf{1}_j^T + \sigma^2 \omega^{-2} \mathbf{I}_j)^{-1} \mathbf{x}_j^* \\
&= \sigma^{-2} \sum_{j=1}^J \mathbf{x}_j^{*T} [\mathbf{I}_j - \mathbf{1}_j (\mathbf{1}_j^T \mathbf{1}_j + \sigma^2 \omega^{-2})^{-1} \mathbf{1}_j^T] \mathbf{x}_j^* \\
&= \sigma^{-2} \sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{x}_j^* - \sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{1}_j (\mathbf{1}_j^T \mathbf{1}_j + \sigma^2 \omega^{-2})^{-1} \mathbf{1}_j^T \mathbf{x}_j^* \\
&= \sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{x}_j^*.
\end{aligned} \tag{5.15}$$

Using a similar argument,

$$\sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{V}_{2,j}^{-1} \mathbf{Y}_j = \sum_{r=1}^R \mathbf{x}_j^{*T} \mathbf{Y}_j \tag{5.16}$$

With these results in mind, we can see that, using the centered value of $\bar{x}_j = \sum_{i=1}^{n_j} x_{ij} / n_j$ as a predictor,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{x}_j^* \right)^{-1} \sum_{j=1}^J \mathbf{x}_j^{*T} \mathbf{Y}_j, \quad \text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\sum_{r=1}^R \mathbf{x}_r^{*T} \mathbf{x}_r^* \right)^{-1}, \tag{5.17}$$

the OLS estimator.

Three-level model. The three-level model is a special case of (5.1) with $L=3$, $r=k$, re \mathbf{Y}_k is the n_k by 1 vector of outcomes within level-3 unit k having elements $\{y_{ijk}\}$. The covariance matrix is $\text{Var}(\mathbf{Y}_k | \mathbf{b}_k) = \mathbf{V}_{3,k} = \omega^2 \mathbf{1}_k \mathbf{1}_k^T + \mathbf{V}_{2,k}$, with $\mathbf{V}_{2,k} = \bigoplus_{j=1}^{J_k} \mathbf{V}_{2,jk}$, and $\mathbf{V}_{2,jk} = \tau^2 \mathbf{1}_{jk} \mathbf{1}_{jk}^T + \sigma^2 \mathbf{I}_{jk}$ where σ^2 is the level-1 variance, τ^2 is the level-2 variance, $\mathbf{1}_{jk}$ is an n_{jk} by 1 vector with all elements equal to unity, and \mathbf{I}_{jk} is the n_{jk} by n_{jk} identity matrix. An example involves students $i=1, \dots, n_{jk}$ nested within classrooms $j=1, \dots, J_k$ that are, in turn, nested within schools $k=1, \dots, K$.

In the case of ω^2, τ^2 , and σ^2 known, the maximum likelihood estimator of the regression coefficients is

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= \left(\sum_{k=1}^K \mathbf{x}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{x}_k \right)^{-1} \sum_{k=1}^K \mathbf{x}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{Y}_k \\
&= \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^K \mathbf{1}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{1}_k & \sum_{k=1}^K \mathbf{1}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{x}_k \\ \sum_{k=1}^K \mathbf{x}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{1}_k & \sum_{k=1}^K \mathbf{x}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{x}_k \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^K \mathbf{x}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{Y}_k \\ \sum_{k=1}^K \mathbf{x}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{Y}_k \end{bmatrix} \quad (5.18)
\end{aligned}$$

We now set

$$\mathbf{x}_k^* = \mathbf{x}_k - \mathbf{1}_k \bar{\mathbf{x}}_k \quad (5.19)$$

where

$$\bar{\mathbf{x}}_k = (\mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{1}_k)^{-1} \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{x}_k = \frac{\sum_{j=1}^{J_k} (\tau^2 + \sigma^2 / n_{jk})^{-1} \bar{x}_{jk}}{\sum_{j=1}^{J_k} (\tau^2 + \sigma^2 / n_{jk})^{-1}}. \quad (5.20)$$

We then have

$$\mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* = \mathbf{0}. \quad (5.21)$$

Substituting \mathbf{x}_j^* for \mathbf{x}_j in (5.18), we now find that

$$\begin{aligned}
\sum_{k=1}^K \mathbf{1}_k^T \mathbf{V}_{3,k}^{-1} \mathbf{x}_k &= \sum_{k=1}^K \mathbf{1}_k^T (\omega^2 \mathbf{1}_k \mathbf{1}_k^T + \mathbf{V}_{2,k})^{-1} \mathbf{x}_k^* \\
&= \sum_{k=1}^K \mathbf{1}_k^T [\mathbf{V}_{2,k}^{-1} - \mathbf{V}_{2,k}^{-1} \mathbf{1}_k (\mathbf{1}_k^T \mathbf{1}_k + \omega^{-2} \mathbf{I}_k) \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1}] \mathbf{x}_k^* \\
&= \sum_{k=1}^K \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* - \sum_{k=1}^K \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{1}_k (\mathbf{1}_k^T \mathbf{1}_k + \sigma^2 \omega^{-2}) \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* \\
&= \mathbf{0} \quad (5.22)
\end{aligned}$$

$$\begin{aligned}
\sum_{k=1}^J \mathbf{x}_k^{*T} \mathbf{V}_{3,k}^{-1} \mathbf{x}_k^* &= \sum_{k=1}^K \mathbf{x}_k^{*T} [\mathbf{V}_{2,k}^{-1} - \mathbf{V}_{2,k}^{-1} \mathbf{1}_k (\mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{1}_k + \omega^{-2} \mathbf{I}_k) \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1}] \mathbf{x}_k^* \\
&= \sum_{k=1}^K \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* - \sum_{k=1}^K \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{1}_k (\mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{1}_k + \omega^{-2} \mathbf{I}_k) \mathbf{1}_k^T \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* \\
&= \sum_{k=1}^K \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* \quad (5.23)
\end{aligned}$$

Using a similar argument,

$$\sum_{k=1}^J \mathbf{x}_k^{*T} \mathbf{V}_{3,k}^{-1} \mathbf{Y}_k = \sum_{k=1}^K \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{Y}_k \quad (5.24)$$

With these results in mind, we can see that, using the centered value of

$$\bar{\mathbf{x}}_k = \frac{\sum_{j=1}^{J_k} (\tau^2 + \sigma^2 / n_{jk})^{-1} \bar{x}_{jk}}{\sum_{j=1}^{J_k} (\tau^2 + \sigma^2 / n_{jk})^{-1}}$$
 as a predictor,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k=1}^J \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* \right)^{-1} \sum_{k=1}^K \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{Y}_k \quad \text{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{k=1}^J \mathbf{x}_k^{*T} \mathbf{V}_{2,k}^{-1} \mathbf{x}_k^* \right)^{-1}, \quad (5.25)$$

Which has the form of a generalized least squares estimator based on a two-level hierarchical linear model.

5.2 Two-Dimensional Confounding

We now consider the case in which observations are nested within the cells of a two-way cross classification. The idea is to remove the confounding associated within levels of across two dimensions. For example, we might have repeated observations on students cross-classified by classrooms. Alternatively, the time-series may be cross classified by students and schools where there are, in addition, classrooms nested within school c . Our model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{R}\mathbf{u} + \mathbf{C}\mathbf{v} + \mathbf{e}, \quad (5.26)$$

$$\mathbf{u} \sim N(\mathbf{0}, \omega^2 \mathbf{I}), \quad \mathbf{c} \sim N(\mathbf{0}, \psi^2 \mathbf{I}), \quad \mathbf{e} \sim N(\mathbf{0}, \mathbf{V}^*).$$

Here \mathbf{R} and \mathbf{C} are matrices of indicators that assign random effects \mathbf{u} to the appropriate “rows” (e.g., children) and \mathbf{v} to the appropriate “columns” (e.g., schools) respectively. Equation (5.26) is a special case of the general model (5.1) with

$$\mathbf{A} = (\mathbf{R} \ \mathbf{C}), \mathbf{b} = (\mathbf{u}^T \ \mathbf{v}^T)^T \text{ and}$$

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega^2 \mathbf{I}, & \mathbf{0} \\ \mathbf{0} & \psi^2 \mathbf{I} \end{bmatrix}. \quad (5.27)$$

Recall from Section 4 that the correct centering requires two conditions:

$$\mathbf{1}^T \mathbf{V}^{*-1} \mathbf{x}^* = \mathbf{0} \quad (5.28)$$

$$\mathbf{x}^{*T} \mathbf{V}^{*-1} \mathbf{A} = (\mathbf{x}^{*T} \mathbf{V}^{*-1} \mathbf{R} \quad \mathbf{x}^{*T} \mathbf{V}^{*-1} \mathbf{C}) = (\mathbf{0} \ \mathbf{0}).$$

This suggests that we regress \mathbf{x} on \mathbf{C} and \mathbf{R} , using generalized least squares with weight matrix \mathbf{V}^{*-1} , then extract residuals \mathbf{x}^* . We illustrate this approach in the case of $\mathbf{V}^* = \sigma^2 \mathbf{I}$.

Illustrative example. Marshall Jean at the University of Chicago has assembled a data set on more than two hundred thousand students moving across more than 500 schools in Chicago with the aim of estimating the impact of certain school-level characteristics, which, along with a vector of time-varying covariates, are collected in the matrix \mathbf{x} . The two-dimensional fixed effects estimation would remove time-varying confounding attributable to students and schools. However, this is a computationally difficult task and imposes limits on modeling that are relaxed when using adaptive centering with random effects, as discussed in Section 3. Can the adaptive centering approach be feasibly implemented in this case?

In principle, we might regress \mathbf{x} on \mathbf{C} and \mathbf{R} , using ordinary least squares (given the assumption $\mathbf{V}^* = \sigma^2 \mathbf{I}$). We would then extract residuals \mathbf{x}^* , achieving condition (5.28). But this is computationally demanding given the dimension of \mathbf{R} (over 200,000 rows). We used the following procedure:

Step 1. Regress \mathbf{x} on \mathbf{R} , save the residuals. This is equivalent to centering around the child mean. Specifically, define \mathbf{x}_{tik} as the vector of explanatory variables for student i attending school k at time t , $t=1, \dots, T_{ik}$; $i=1, \dots, n$; $k=1, \dots, K$. Then we have $\mathbf{x}_{tik}^* = \mathbf{x}_{tik} - \bar{\mathbf{x}}_{\cdot i}$.

Step 2. For each observation, regress \mathbf{C} on \mathbf{R} , save the residuals. This is easier than it sounds. Simply define dummy variable $C_{tik} = 1$ if student i attends school k at time t ; $C_{tik} = 0$ otherwise. Do this for each school $k=1, \dots, K$ so that there are K dummy variables per occasion per student. Now compute $C_{tik}^* = C_{tik} - n_{ik} / n_i$, where n_{ik} is the number of observations for student i in school k and n_i is the total number of observations for student i . Thus n_{ik} / n_i is the proportion of student i 's observations that occurred while in school k . The collection of those is equivalent to the predicted value of \mathbf{C} given \mathbf{R} so that C_{tik}^* are the residuals.

Step 3. Compute a regression with $\mathbf{x}_{tik}^* = \mathbf{x}_{tik} - \bar{\mathbf{x}}_{\cdot i}$ as the outcome and with J predictors $C_{tik}^* = C_{tik} - n_{ik} / n_i$, $k=1, \dots, K$. Save the residuals from this regression, that is, save $\mathbf{x}_{tik}^{**} = \mathbf{x}_{tik}^* - \hat{E}(\mathbf{x}_{tik}^* | C_{i1}^*, \dots, C_{iK}^*)$. These in fact are the variables to be used in our regressions. This approach satisfies (5.28), removing time-invariant confounding attributable to children and schools. In the special case of balanced data, that is, when each student is observed the same number of times in each school (as in the hypothetical case of Section 3), the result is $\mathbf{x}_{tik}^{**} = \mathbf{x}_{tik} - \bar{\mathbf{x}}_{\cdot i} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}}_{\dots}$ as in Section 3.

References

Green, W., 1997, *Econometric Analysis*, 3rd Ed., Prentice Hall, Saddle River.

Hong, G. and Raudenbush, S.W., (in press) Causal inference for time-varying instructional treatments. *The Journal of Educational and Behavioral Statistics*.

Huber, P.J., (1967) The behavior of maximum likelihood estimates under non-standard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 221-233.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-41.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*. Vol. 29, No. 1, pp 67.101.

Raudenbush, S.W., Bryk, A.S., Cheong, Y., & Congdon, R.T. (2000). *HLM 5: Hierarchical Linear and Nonlinear Modeling*. Chicago: Scientific Software International.

Robins, J. (2000). Marginal structural models versus structural nested models as tools for causal inference, in M. Elizabeth Halloran and Donald Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials*. (pp. 95-134). New York: Springer.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, pp.41-55.

Rubin, D.B., (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81, 961-962.

Table 1.
Outcome data for 20 hypothetical kids by 9 teachers nested with 3 schools

		School 1				School 2				School 3		
	Teacher	1	2	3		4	5	6		7	8	9
	x	-1	0	1		-1	0	1		-1	0	1
w	Child											
0	1			-2.4102				2.4628				6.2245
1	2			3.6396			4.1441					11.0898
1	3		2.1827					10.1339				12.3134
0	4			-3170				3.6596			4.8397	
0	5		-.0727				1.6280				6.0525	
0	6		-2.7852				1.4795					10.0131
0	7		.2350					6.0839			7.5142	
0	8	-.8803				3.5167						9.7337
0	9	-1.5147					5.8636					10.2860
0	10			2.6814				7.6954				10.0192
1	11	4.4966				9.5578				11.1152		
1	12	4.7195				8.2204					14.6855	
1	13	4.3609						12.6474			16.8547	
1	14	4.7778					11.9663					18.3998
1	15		8.5264				12.9066					18.6272
1	16		8.6820			11.8265					17.0661	
1	17		9.5595				13.8078				16.3071	
1	18	5.6075				12.7943						21.075
1	19	8.9094					13.5301					20.049
0	20	6.3465				7.3268				11.5147		

Table 2

One-Dimensional Control: OLS Fixed Child Effects

$$y_{ijk} = \theta + \beta x_j + u_i + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2),$$

$u_i, i = 1, \dots, 19$ fixed

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	t	Sig.
Intercept	13.894087	2.217045	6.267	.000
x	5.498095	.865904	6.350	.000
[childid=1.00]	-17.299841	3.366029	-5.140	.000
[childid=2.00]	-11.268353	3.227033	-3.492	.001
[childid=3.00]	-9.349477	3.227033	-2.897	.006
[childid=4.00]	-14.832045	3.227033	-4.596	.000
[childid=5.00]	-11.358169	3.013434	-3.769	.001
[childid=6.00]	-12.825538	3.108690	-4.126	.000
[childid=7.00]	-11.115732	3.108690	-3.576	.001
[childid=8.00]	-9.770723	3.013434	-3.242	.002
[childid=9.00]	-9.015820	3.013434	-2.992	.005
[childid=10.00]	-12.593491	3.366029	-3.741	.001
[childid=11.00]	-.006149	2.886346	-.002	.998
[childid=12.00]	-1.020260	2.900742	-.352	.727
[childid=13.00]	-.773729	2.943507	-.263	.794
[childid=14.00]	-2.179455	3.013434	-.723	.474
[childid=15.00]	-2.373398	3.108690	-.763	.450
[childid=16.00]	.463474	2.943507	.157	.876
[childid=17.00]	-.669300	3.013434	-.222	.825
[childid=18.00]	1.097582	2.943507	.373	.711
[childid=19.00]	.268870	3.013434	.089	.929
[childid=20.00]	0(a)	0	.	.

Estimates of Covariance Parameters

Parameter	Estimate
σ^2	12.496491

Table 3

**One-Dimensional Control:
Child random effects with person-mean centered x**

$$y_{ijk} = \theta + \beta(x_{tik} - \bar{x}_i) + u_i + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2),$$

$$u_i \sim N(0, \tau^2)$$

Note this gives the same coefficient, standard error, and residual variance estimate as the student fixed effects model.

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.
Intercept	8.029549	.927088	19	8.661	.000
$(x_{tik} - \bar{x}_i)$	5.498095	.865904	39.000	6.350	.000

Estimates of Covariance Parameters

Parameter	Estimate
σ^2	12.496491
τ^2	13.024353

Table 4

Two dimensional controls: OLS fixed child and school effects

$$y_{ijk} = \theta + \beta x_j + u_i + s_k + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2),$$

$u_i, i = 1, \dots, 19$ fixed

$s_k = 1, 2$ fixed

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	T	Sig.
Intercept	14.642231	.630345	37	23.229	.000
X	2.573106	.287937	37	8.936	.000
[childid=1.00]	-	.998365	37	-11.469	.000
[childid=2.00]	11.449864	.946257	37	-6.756	.000
[childid=3.00]	-6.393372	.946257	37	-4.729	.000
[childid=4.00]	-4.474496	.946257	37	-4.729	.000
[childid=5.00]	-9.957064	.946257	37	-10.523	.000
[childid=6.00]	-8.433180	.864876	37	-9.751	.000
[childid=7.00]	-8.925554	.901385	37	-9.902	.000
[childid=8.00]	-7.215747	.901385	37	-8.005	.000
[childid=9.00]	-6.845734	.864876	37	-7.915	.000
[childid=10.00]	-6.090831	.864876	37	-7.042	.000
[childid=11.00]	-6.743514	.998365	37	-6.755	.000
[childid=12.00]	-.006149	.815539	37	-.008	.994
[childid=13.00]	-.045263	.821167	37	-.055	.956
[childid=14.00]	1.176263	.837825	37	1.404	.169
[childid=15.00]	.745534	.864876	37	.862	.394
[childid=16.00]	1.526586	.901385	37	1.694	.099
[childid=17.00]	2.413467	.837825	37	2.881	.007
[childid=18.00]	2.255688	.864876	37	2.608	.013
[childid=19.00]	3.047574	.837825	37	3.637	.001
[childid=20.00]	3.193858	.864876	37	3.693	.001
[childid=20.00]	0(a)	0	.	.	.
[schoolid=1.00]	-7.679293	.367143	37	-20.916	.000
[schoolid=2.00]	-3.340106	.347120	37	-9.622	.000
[schoolid=3.00]	0(a)	0	.	.	.

Estimates of Covariance Parameters

Parameter	Estimate
σ^2	.997655

Table 5

Two-Dimensional Controls: Random child and school effects with interaction-contrast centering

$$y_{tijk} = \theta + \beta(x_{tijk} - \bar{x}_i - \bar{x}_k + \bar{x}) + u_i + s_k + \varepsilon_{tijk},$$

$$\varepsilon_{tijk} \sim N(0, \sigma^2)$$

$$u_i \sim N(0, \tau^2),$$

$$s_k \sim N(0, \psi^2)$$

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	t	Sig.
Intercept	8.029463	2.851520	2.816	.083
$x_{tijk} - \bar{x}_i - \bar{x}_k + \bar{x}$	2.573106	.287937	8.936	.000

Estimates of Covariance Parameters

Parameter	Estimate
σ^2	.997655
τ^2	16.857298
ψ^2	21.815022