Empirical Strategies for Various Manifestations of Multilevel Data

Robert J. Franzese, Jr.
Department of Political Science,
University of Michigan, Ann Arbor, MI
e-mail: franzese@umich.edu

Equivalent separate-subsample (two-step) and pooled-sample (one-step) strategies exist for any multilevel-modeling task, but their relative practicality and efficacy depend on dataset dimensions and properties and researchers' goals. Separate-subsample strategies have difficulties incorporating cross-subsample information, often crucial in time-series cross-section or panel contexts (subsamples small and/or cross-subsample information great) but less relevant in pools of independently random surveys (subsamples large; cross-sample information small). Separate-subsample estimation also complicates retrieval of macro-level-effect estimates, although they remain obtainable and may not be substantively central. Pooled-sample estimation, conversely, struggles with stochastic specifications that differ across levels (e.g., stochastic linear interactions in binary dependent-variable models). Moreover, pooled-sample estimation that models coefficient variation in a theoretically reduced manner rather than allowing each subsample coefficient vector to differ arbitrarily can suffer misspecification ills insofar as this reduced specification is lacking. Often, though, these ills are limited to inefficiencies and standard-error inaccuracies that familiar efficient (e.g., feasible generalized least squares) or consistent-standard-error estimation strategies can satisfactorily redress.

1 Introduction

Multilevel data are lower, micro-level data nested within higher, macro-level units. Political science examples include survey respondents nested within countries or states, elections nested within countries, time periods nested with nations or nation dyads, and many more. The levels may exceed two, such as survey respondents within elections within countries, voters within districts within countries, time periods within directed-dyads within dyads, etc. *Level 1* or the *micro-level* is the lowest level or smallest unit of analysis; higher levels are *Level 2*, *Level 3*, etc., or *macro-level(s)*. Common multilevel datasets in political science include cross-context surveys, which contributions to this volume analyze; panel (survey) data, containing repeated surveys of the same individuals; time-series cross-section (TSCS) datasets common in comparative/international politics and political economy, which typically nest time periods within countries; and datasets

---

commonly used in international relations, which often nest time periods within dyads or directed dyads.[1]

In considering how to analyze such data empirically effectively for various goals, one must first recognize that typical dataset dimensions (i.e., numbers of micro- and macro-level observations) vary across dataset types and substantive contexts, as do plausible variance-covariance structures for variables, errors, and outcomes (i.e., systematic, stochastic, and total components). Practical and effective empirical-modeling strategies vary accordingly with these dimensions and properties and with researchers' goals and questions. Strategies that make great sense with large numbers of micro-level observations that vary independently across smaller numbers of contexts, as is common when pooling independently random surveys across countries, would not necessarily be as sensible when pooling observations related over time and across contexts, as is typical in international relations and political economy applications. There, with moderate numbers of both micro- and macro-level units and/or with observations related across contexts, alternative strategies become necessary or more practical or effective. In all cases, analysts will want to keep their methods as simple, powerful, and accurate as possible and to keep their own research questions and goals always central to their methodological choices, but what this implies practically may differ across research contexts.

In principle, two-step, separate-subsample estimation can achieve anything achievable in one-step, pooled-sample estimation and vice versa, but which strategy will prove more practical or effective depends on dataset dimensions and properties and on substantive contexts and goals. In general, two-step strategies have difficulties incorporating cross-subsample information, such as when variance-covariance or coefficient parameters may be equal, proportional, or otherwise related or when micro-level outcomes are interdependent, across contexts. Such information often exists and is sometimes substantively central in comparative and international politics and political economy; furthermore, such information is often indispensable there regardless of its substantive centrality given the practical limitations set by typical dataset dimensions. On the other hand, cross-subsample information is usually absent or small, and anyway less essential, in pools of large, independently random surveys, although ignoring some kinds of cross-sample dependence can induce biases.

Two-step estimation also tends to obscure and complicate, although certainly not to debar, the retrieval of estimates of the effects of macro-level variables as opposed to those of micro-level variables or of micro-level variables as conditioned by macro-level ones. The separate-sample, micro-level equations produce the micro-level effects directly, and how that effect depends on macro-level factors emerges directly from a second, macro-level estimation. To obtain the macro-level effect on the outcome, however, would require system-of-equations estimation in that second stage. Micro-behavioral researchers may be less interested in these macro-level effects, per se, but institutionalists and political economists would often find them equally central to their interests.

Pooled-sample estimation, conversely, produces all three effect estimates directly and renders leveraging of cross-sample information simple. However, maintaining efficiency, accurate standard-error estimation, and, in some cases, even unbiasedness and consistency can require additional care. One-step strategies become progressively more complicated as

---

[1]Concrete examples of each include, respectively, *Comparative Study of Electoral Systems*, *National Election Studies*, and *World Values Surveys*; NES panel studies; IMF, World Bank, or OECD datasets; *Comparative Manifestos*, which pools parties' election manifestos by election across several democracies; and political-institutional datasets like *Polity* or *Freedom House*; and the *Correlates of War*, *Militarized Interstate Disputes* datasets.

stochastic complexities like unit-specific covariances and stochastic interactive effects arise, and they have particular difficulties with stochastic-model specifications that differ across levels. For example, micro-level effects on binary outcomes being linear-interactively conditioned stochastically by macro-level factors would require nesting linear-normal likelihoods within binomial likelihoods: feasible, but not so simple. More-over, pooled-sample estimation that models cross-subsample coefficient variation rather than allowing each subsample coefficient vector to differ arbitrarily can suffer various misspecification ills insofar as the theoretically reduced model of context conditionality is lacking. In many cases, though, these ills will be limited to inefficiency and standard-error inaccuracy, which can be satisfactorily redressed by familiar efficient (e.g., feasible generalized least squares, or FGLS) or consistent-standard-error estimation strategies.

Elaboration and discussion of these points unfolds thus. Section 2 introduces a generic multilevel model and the three generic substantive research questions that researchers use such models to evaluate empirically: effects of micro-level and macro-level characteristics and how each depends on the other. As argued more fully there, if "good" estimation (unbiased, consistent, efficient, plus simple and presentationally effective) of these effects and their variance-covariance (standard-errors)[2] are the goals, then one wants separate-subsample estimates per se only to satisfy intrinsic interest in unique models for each subsample, for sensitivity analysis of whether restrictions implied by pooled and reduced models hold across subsamples, or, relatedly, if pooled-sample coefficient or standard-error estimation would lack "good" properties. Therefore, Section 3 begins by discussing the sample and theoretical/substantive conditions under which the simplest possible estimator—full pooling with common parameters across samples—would suffer no ills, and proceeds to complicate the true stochastic and systematic environment from there. The concluding Section 4 contrasts typical panel and TSCS data in comparative and international politics and political economy with that of cross-context compilations of surveys in comparative micro-behavioral research based on the preceding discussion and existing simulation studies of alternative estimation strategies.

2   A Generic Multilevel Model and Typical Multilevel Research Questions

Multilevel or hierarchical models are any $y_{ij} = f(\mathbf{X}_{ij}, \varepsilon_{ij})$, i.e., any model in which outcomes, explanatory factors, and/or stochastic components occur at nested, micro and macro levels, i and j,[3] but, of course, they become more interesting when some arguments vary only at macro levels, $\mathbf{Z}_j$, and others vary at micro levels, $\mathbf{X}_{ij}$, and especially when interactions occur across levels, $\mathbf{X}_{ij}\mathbf{Z}_j$, and/or when the stochastic properties of ε_{ij} pattern by level. Follow Bowers and Drake (2005) to consider this general expression of a hierarchical linear model (HLM), with one z_j and x_{ij}:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}; \tag{1}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_j + u_{0j}; \tag{2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}z_j + u_{1j}; \tag{3}$$

$$\Rightarrow y_{ij} = \gamma_{00} + \gamma_{01}z_j + u_{0j} + (\gamma_{10} + \gamma_{11}z_j + u_{1j})x_{ij} + \varepsilon_{ij}; \tag{4}$$

$$\Rightarrow y_{ij} = \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}z_jx_{ij} + (u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}). \tag{5}$$

---

[2]The bias, consistency, and efficiency of *standard errors* refer to their properties relative to the true *standard deviation* of parameter estimates across repeated samples under the model assumptions.

[3]More generally, the number of levels can exceed two, but we will consider only two to keep discussion simple.

Equation (1) gives a bivariate linear-regression model of outcome, y_{ij}, as linear-additive function of explanator, x_{ij}, and additively separable stochastic component, ε_{ij}. Equation (2) adds complications that another explanator, z_j, which varies only across and not within macro level, j, also affects y_{ij} and that it does so with some error, u_{0j}, which also varies only across macro levels, j. At this point, we have a (trivariate) random-effects model with two explanators, x_{ij} and z_j, and a compound error term, $u_{0j} + \varepsilon_{ij}$, with u_{0j} being the macro-unit-specific random effect. Equation (3) adds further complications that x_{ij} and z_j interact in determining the outcome, y_{ij}, implying that the effect of x_{ij} depends on z_j and, vice versa, that the effect of z_j depends on x_{ij}, and that these conditioning effects, too, occur with macro-unit-specific error, u_{1j}. At this point, as seen in Eq. (4), we have a generic HLM, which also happens to be a special, limited type of random-effects and random-coefficients model (as explained below). As seen in Eq. (5), however, the generic HLM is also quite similar to the familiar linear-interactive model (Franzese and Kam 2005; Brambor et al. 2005), with explanators x_{ij}, z_j, and $x_{ij}z_j$, except that the HLM possesses a compound error term, $u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}$, which complicates matters.

This similarity of HLM to simple linear-interaction models suggests our central question: Under what conditions will simple linear-interactive regression models reflecting theoretical propositions about micro, macro, and micro-macro-interactive effects suffice, and under what conditions will multiple stages of estimation or complicated HLM or other elaborate econometric strategies be preferable? As already outlined, the answer seems to depend on sample dimensions and properties, the availability and importance of cross-sample information, and the researcher's goals.

For concreteness, consider a comparative political economy and a comparative micro-behavioral example in which the outcomes, y_{ij}, are, respectively, fiscal policy stance (budget balance) at time i in country j and a feeling-thermometer score for a center-right party of respondent i in country j. The micro-level explanators, x_{ij}, could be, e.g., government partisanship at time i and the income level of respondent i in country j. The macro-level factors, z_j, could be, respectively, the district magnitude of the (assumed time-invariant) electoral system and income inequality in j. A researcher hypothesizes, in case one, that policy makers elected in larger-district-magnitude systems weigh public goods and broad redistribution more heavily relative to narrowly targeted distribution (i.e., pork) than do those elected in smaller district-magnitude systems. Fiscal activism as gauged by budget balances will reflect redistributive and public good (Keynesian macro-management) efforts more than distributive ones, so larger district magnitudes, z_j, should increase deficits (reduce y_{ij}). Left governments, too, x_{ij}, might have this effect, and especially when elected in systems of larger district magnitude; thus the interactive term $x_{ij}z_j$ also enters. A case two researcher expects poorer respondents, x_{ij}, to feel cooler toward center-right parties, especially as income is more unequally distributed in their country (thus the interactive term $x_{ij}z_j$), and all respondents in more-unequal countries might feel less warmly toward center-right *parties ceteris paribus* compared to those in more equal countries (thus z_j).

Generically, then, researchers seek to estimate three kinds of effects in multilevel models: effects on y_{ij} (deficits, thermometer) of micro-level factors (government partisanship or respondent income), x_{ij}, i.e., $\frac{\partial y}{\partial x}$, which may vary across macro-level contexts, j, depending on macro-level factors (district magnitude or inequality), z_j; the effects of macro-level factors (magnitude or inequality), z_j, on y_{ij}, i.e., $\frac{\partial y}{\partial z}$, which may vary across micro-level units, ij, depending on micro-level factors (partisanship or income), x_{ij}; and if and how the effects of these micro- and macro-level factors depend on the other

variable, i.e., $\frac{\partial^2 y}{\partial x \partial z} \equiv \frac{\partial^2 y}{\partial z \partial x}$.[4] Mathematically in model (5), these effects (in expectation) are, respectively:

$$E\left(\frac{\partial y_{ij}}{\partial x_{ij}}\right) = E(\beta_{1j}) = E(\gamma_{10} + \gamma_{11}z_j + u_{1j}) = \gamma_{10} + \gamma_{11}z_j; \tag{6}$$

$$E\left(\frac{\partial y_{ij}}{\partial z_j}\right) = \gamma_{01} + \gamma_{11}x_{ij}; \tag{7}$$

$$E\left(\frac{\partial^2 y_{ij}}{\partial x_{ij}\partial z_j}\right) \equiv E\left(\frac{\partial^2 y_{ij}}{\partial z_j \partial x_{ij}}\right) = \gamma_{11}. \tag{8}$$

Equations (6)–(8) underscore a certain asymmetry in the random coefficients of the generic HLM to which we alluded above. Whereas micro-level effects, Eq. (6), vary *stochastically* across macro-level units, macro-level effects, Eq. (7), vary *nonstochastically* across micro-level units, and the interactive effect, Eq. (8), is *constant*. These features reflect an assumption that macro-, contextual-, or country-level error components arise in the constant and in the coefficient on x_{ij}, i.e., in Eqs. (2) and (3), but that such error components do not arise elsewhere in the model. Each observation i in unit j experiences the same realization of the macro-unit-specific errors, one additively, u_{0j}, and one, u_{1j}, in its coefficient on x_{ij}, β_{1j}. A fuller random-effects and coefficients model would decompose γ_{01} further into a linear-additive function of x_{ij} and an error term, u_{2ij}, and perhaps add a fourth error component, u_{3ij}, to the interaction parameter, γ_{11}, as well. We will follow standard HLM practice, but perhaps the reader can infer the implications for the more general random-coefficients model by analogy to the discussion of the particular HLM form considered here (i.e., u_{0j} in β_{0j} and u_{1j} in β_{1j} only).

Notice, finally, that insofar as researchers aim to estimate Eqs. (6)–(8), the effects of x_{ij} and of z_j on y_{ij}, and how these depend on each other—e.g., how partisanship and districting (inequality and income) interact to explain fiscal policy (party affinity)—they are actually *uninterested* in β_{0j} and β_{1j} per se. That is, for these interests, unique estimates of intercepts, β_{0j}, and effects of x on y, β_{1j}, in each macro-unit j, are not the goal; the goal is to estimate the *systematic* or *explicable* aspects of effects of x and z and how each depends on each other, i.e., γ_{01}, γ_{10}, and γ_{11} (and perhaps the conditional mean, γ_{00}, too). Put differently, we do not seek a specific model for each j, but a model of the outcome, y_{ij}: not separate models of French, German, Japanese, and every other j's politics/political economy, not unique models of x-y relations for each election j, but a model of comparative politics/political economy or of electoral politics. If β_{0j} and β_{1j} vary across contexts j, comparative researchers seek to model, understand, and explain this interesting phenomenon by variations in contextual factors, z_j.

Interest in estimating β_{0j} and β_{1j} directly, therefore, arises only (a) to satisfy any intrinsic interest in unique models for each subsample, (b) for sensitivity analysis of whether restrictions implied by a pooled and reduced model seem to hold across subsamples, or (c), relatedly, if one-step reduced-form coefficient or standard-error estimation lacks *good* properties. Apart from inherent curiosity in specific models for each j, researchers will want to estimate j context-unique models only if going directly to theoretically reduced models might mislead. That is, variation in effects across contexts is to be explained, not merely

---

[4]Note that z modifies the effect of x on y identically to how x modifies the effect of z on y; these statements are logically (and so mathematically) identical. Likewise, x and z each has only additive effect on y if the interaction does not exist.

described, so what remains arbitrarily (i.e., inexplicably) variant across j matters only insofar as giving it insufficient attention would harm coefficient or variance-covariance estimates of the theoretically interesting model. Accordingly, the next section considers estimation strategies for multilevel data from the perspective of asking under what conditions might one wish to estimate anything other than Eq. (5) directly in one simple step: pooled linear-interactive ordinary least squares (OLS).

3 Estimation Strategies

3.1 *Fully Pooled, Context-Unconditional OLS*

The first, and simplest, possible strategy would be to estimate Eq. (1), i.e., to regress y_{ij} on x_{ij}, by fully pooled OLS:

$$y_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{\varepsilon}_{ij} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} \quad where: \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}; \ \hat{V}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (9)$$

By Gauss-Markov, these are BLUE (best, i.e., minimum-variance, linear unbiased estimates) *iff* $\gamma_{01} = u_{0j} = \gamma_{11} = u_{1j} = 0$ and $V(\varepsilon_{ij}) = \sigma^2$. That is, fully pooled OLS with just the micro-level regressor is optimal if and only if the effect of x on y is constant and nonstochastic across contexts (i.e., $\gamma_{01} = \gamma_{11} = 0$, so Z_j does not matter, and $u_{0j} = u_{1j} = 0$) and outcomes are homoskedastic across and within contexts (i.e., constant variance and no correlation: $V(\varepsilon) = \sigma^2\mathbf{I}$). Obviously, this is the least interesting, and usually quite implausible, case. Notice, however, that if Z_j truly does not matter ($\gamma_{01} = \gamma_{11} = 0$), then even if $u_{0j} \neq 0$ or $u_{1j} \neq 0$, i.e., even if macro-unit specific error components (random effects/coefficients) exist or if the stochastic component is nonspherical (heteroskedastic/correlated errors), then this starkest fully pooled OLS would nonetheless produce unbiased and consistent coefficient estimates provided $E(u_{0j}, u_{0j}, \varepsilon_{ij} \mid X) = 0$ (i.e., the usual Gauss-Markov requirement that regressors and residuals not covary):

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) = E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\{\gamma_{00} + \gamma_{10}x_{ij} + (u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij})\} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\gamma_{00} + \gamma_{10}x_{ij} + E\{u_{1j}x_{ij}\} + E\{u_{0j}\} + E\{\varepsilon_{ij}\}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\gamma_{00} + \gamma_{10}x_{ij} + E\{u_{1j}\}x_{ij} + 0 + 0] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\gamma_{00} + \gamma_{10}x_{ij} + 0 \cdot x_{ij}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\gamma_{00} + \gamma_{10}x_{ij}] = [\gamma_{00}\gamma_{10}]'. \quad (10)
\end{aligned}
$$

Thus, even with hierarchically patterned error components, OLS coefficient estimates are unbiased and consistent unless any stochastic component correlates with regressors. OLS coefficient estimates are inefficient, however, and OLS variance-covariance estimates (i.e., standard errors) are wrong (biased, inconsistent, inefficient). In other words, macro-unit-specific error components (i.e., random coefficients/effects: stochastic variation across j in effects of x or z) only harm OLS efficiency and standard-error estimation; they do not induce bias or inconsistency *in the linear-regression model*.[5]

---

[5]More precisely, the following discussion applies to models with additively separable stochastic components, such as linear-regression models, but not necessarily models with nonseparable stochastic components, such as logit or probit.

OLS is inefficient and OLS standard errors are incorrect because OLS ignores the nonconstant error variance and correlation that the macro-unit-specific error components induce:[6]

$$
\begin{aligned}
\hat{V}(\hat{\boldsymbol{\beta}}) = V\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[V\left\{\gamma_{00} + \gamma_{10}x_{ij} + (u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij})\right\}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[V\left\{u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}\right\}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[V\left\{u_{1j}x_{ij}\right\} + V\left\{u_{0j}\right\} + V\left\{\varepsilon_{ij}\right\}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[V\left\{u_{1j}\right\}\cdot x_{ij}^2 + V\left\{u_{0j}\right\} + V\left\{\varepsilon_{ij}\right\}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}
$$

$$(11)$$

Even if each error component, u_{0j}, u_{1j}, and ε_{ij}, is spherical (constant variance and uncorrelated), the term in square brackets (i.e., variance of y_{ij}) will not reduce to $\sigma^2 I$ because macro-unit-specific error components, u_{0j} and u_{1j}, differ across j and because u_{1j} multiplies a variable, x_{ij}, and so varies. Thus Eq. (11) does not reduce to the OLS standard-error formula, $\sigma^2(X'X)^{-1}$. To enhance efficiency and obtain accurate standard errors by FGLS, however, one need only estimate the induced heteroskedasticity. As seen from the term in square brackets, one could do this simply by regressing squared estimated residuals on macro-unit indicators and those indicators times x^2, plus whatever patterns one expects in ε_{ij}.[7] As Beck and Katz (1995, 1996) famously showed, however, whether FGLS enhances efficiency and improves standard-error estimation *truly*, and not merely *apparently*, depends on how many parameters (relative to observations) one must estimate in this step, which degrees-of-freedom consumption FGLS will ignore in its next step.[8]

Rendering OLS standard-error estimates consistent ("robust") to the induced heteroskedasticity is even simpler. As Eq. (11) shows, as always, OLS standard errors are inaccurate only insofar as the expression between the $(X'X)^{-1}$ terms differs from $(X'X)$ times a constant (σ^2). That is, they are unbiased (and consistent) if the term in square brackets, the nonsphericity of the error-variance, is unrelated on average (and in the limit) to the x's, x^2's, and x cross products contained in the X' and X pre- and post-multiplying that term. As the first term in square brackets reveals, random effects *do* generally imply biased and inconsistent OLS standard errors because the induced nonsphericity *is* related to x. The macro-level-specific error components plausible in multilevel data, being random effects, have this implication, and also a nonconstant variance by unit, or *clustering*, as seen in that first term and the second term. If the second term correlates with the x's, x^2's, or x cross products, then it adds further bias to OLS standard-error estimates; otherwise they induce "only" further inefficiency.

To render the standard-error estimates consistent, then, one must use a formula that retains the $X'[\cdot]X$ expression in a form capturing the clustering pattern and/or heteroskedasticity. For example, for simple random effects/coefficients without clustering

---

[6]The step from the second to the third line of Eq. (11) assumes that the error components are uncorrelated. If correlated, this and the next expression would include (two times) the additional covariance terms. Furthermore, the assumption that all error components are uncorrelated with the regressors must be maintained here as in all other estimators considered.

[7]This auxiliary regression should include x_{ij} also if the macro-unit-specific shocks are correlated, and it should include all x_{ij} involved in stochastic interactions in the case of multiple such interactions.

[8]Here, extra parameters in FGLS are few (see preceding text and note 7), so the issue is likely small.

(i.e., where the error components in Eqs. (2) and (3) are not macro-unit-specific and shared across micro units within cluster but specific to each observation), White's familiar heteroskedasticity-consistent standard errors will suffice:

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \frac{1}{N-k}(\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{i=1}^{N}(e_i\mathbf{x}_i)(e_i\mathbf{x}_i)'\right](\mathbf{X}'\mathbf{X})^{-1}$$
$$= \frac{1}{N-k}(\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{i=1}^{N}e_i^2\mathbf{x}_i\mathbf{x}_i'\right](\mathbf{X}'\mathbf{X})^{-1},$$

(12)

where i indexes observations across all levels, $\mathbf{x}_i$ is the ith vector of observations on the k regressors, and N is the number of observations.[9] Since this formula accounts movements in variance (or, rather, e_i^2) that relate to $\mathbf{x}_i\mathbf{x}_i'$, i.e. to the x's, x^2's, and x cross products, it is robust to the heteroskedasticity induced by random effects, which is also precisely the sort that biases OLS standard errors. As the small-sample correction of $N/(N-k)$ sometimes multiplied to Eq. (12) (see note 9) suggests, such heteroskedasticity-*consistent* standard errors have proven in Monte Carlo simulations (as reported in Greene 2003, p. 220) to have reasonable properties even in fairly *limited* samples; the small-sample correction, for example, suggests bias of about 10% for $N = 55$, $k = 5$.[10]

For the *clustering* heteroskedasticity induced by the error structure expected in multilevel data, a consistent standard-error estimate must account the common error components within macro units:

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \frac{1}{N-k}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{j=1}^{n_c}\left\{\left[\sum_{i=1}^{n_j}e_i\mathbf{x}_i\right]\left[\sum_{i=1}^{n_j}e_i\mathbf{x}_i\right]'\right\}\right)(\mathbf{X}'\mathbf{X})^{-1},$$

(13)

where n_j is the number of observations i in macro-level (cluster) j, and n_c is the number of clusters.[11] The formula again adjusts for patterns in squared residuals related to $\mathbf{x}_i\mathbf{x}_i'$ but now sums these cross-products of regressors with squared residuals first within the macro level, then sums those sums across clusters, thereby allowing the nonsphericity a pattern clustered by j. Again, such clustered standard errors are *consistent* to exactly the sort of nonspherical error structures that multilevel data are expected to induce. Good *limited-sample* properties, however, now require large numbers of macro-level units as well as observations-to-regressors ratios. The small-sample adjustment suggested here (see note 11), $[N/(N-k)][J/(J-1)]$, gives corrections of about +17% for, say, 100 observations in 20 macro-level units (5 per unit) with 10 regressors, but this drops to about +11% if observations *per* unit, I, double to 10, to about 8% if the number of units, J, doubles, or to about +5% if both I and J double.[12]

---

[9]Davidson and MacKinnon (1993, p. 554) strongly suggest a finite-sample correction of replacing e_i^2 by $e_i^2/(1 - \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i)$, which scales estimated squared residuals by their variance, or of multiplying Eq. (12) by $N/(N-k)$, which inflates estimates by a factor reflecting the number of regressors as a percentage of degrees of freedom. Accumulating simulation work favors their suggestion.

[10]In *Stata*, obtaining these estimates is as simple as typing ", *robust*" or ", *r*" at the end of a line.

[11]As with pure heteroskedasticity (see note 9), a finite-sample (degrees-of-freedom) correction, $[n_c/(n_c - 1)]$ $[(N - 1)/(N - k)]$, is suggested. This inflates standard errors as there, but now multiplicatively further, by a declining function of J. Again, simulations strongly support using such adjustments.

[12]As before, obtaining these estimates in *Stata* is simple; one types ", *cluster(J-indicator-name)*" at the end of a line.

3.2 *Fully-Pooled, Linear-Interactive OLS*

Of course, if Z_j matters, i.e., if $\gamma_{01} \neq 0$ or $\gamma_{11} \neq 0$ such that outcomes exhibit systematic and/or stochastic variation across contexts, which, after all, is what interested us in multilevel models *ab initio*, then estimating context-conditional Eq. (5) (with $k = 4$) by context-unconditional Eq. (9) (with $k = 2$) suffers omitted-variable bias. Researchers interested in context conditionality, which necessarily includes multilevel modelers, must include z_j and/or $x_{ij}z_j$.

Consider, then, estimating the full model of Eq. (5) by OLS:

$$y_{ij} = \hat{\gamma}_{00} + \hat{\gamma}_{10}z_j + \hat{\gamma}_{01}x_{ij} + \hat{\gamma}_{11}x_{ij}z_j + \hat{\eta}_{ij} \quad where \ \ \eta_{ij} = u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}. \quad (14)$$

Again by Gauss-Markov, this fully pooled linear-interactive OLS regression is BLUE *iff* $u_{0j} = u_{1j} = 0$ and $V(\varepsilon_{ij}) = \sigma^2$. That is, as usual, OLS requires a spherical stochastic component, here the compound η_{ij}, for efficiency and accurate standard-error estimation, and it requires this residual to be uncorrelated with regressors for unbiasedness and consistency. Again, provided macro-unit-specific components, u_{0j} and u_{1j}, which can now represent the portion of cross-contextual variation not or insufficiently modeled theoretically by z_j and $x_{ij}z_j$, are uncorrelated with the regressors,[13] OLS coefficient estimates remain unbiased and consistent, although inefficient, but its reported standard errors are inaccurate. Again, the inefficiency and standard-error inaccuracy arise even if u_{0j} and u_{1j} are spherical:

$$\hat{V}(\hat{\beta}) = V\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right\}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[V\left\{\gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}z_j + (u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij})\right\}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[V\left\{u_{1j}\right\} \cdot x_{ij}^2 + V\left\{u_{0j}\right\} + V\left\{\varepsilon_{ij}\right\}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \quad (15)$$

The $\gamma_{01}z_j$ and $\gamma_{11}x_{ij}z_j$ terms are the only differences from Eq. (11). Being non stochastic, they vanish from the last line, so the upshot is identical. Again, if desired, one enhances efficiency via FGLS by the same mechanics as before, and heteroskedasticity-consistent Eq. (12) or cluster-consistent Eq. (13) will render standard-errors *consistent* to, respectively, pure random-effects/coefficients or clustered stochastic components, with the same small-sample concerns as before.

Therefore, if researchers aim to estimate the effects of micro- and macro-level factors and their interactions (e.g., interactive effects of partisanship and magnitude on fiscal policy or of income and inequality on party affinity), little argument has yet arisen against estimating pooled OLS models specified to reflect those interactive propositions. OLS offers unbiased and consistent, although inefficient, coefficient estimates. If sample degrees of freedom are favorable, FGLS could enhance efficiency and is straightforward to implement. OLS, or FGLS that incompletely models the induced nonsphericity, yields biased and inconsistent standard errors, but appropriate robust estimators easily render these heteroskedasticity or cluster consistent. Under what conditions, then, would one consider alternative separate-subsample or HLM estimation strategies?

---

[13]The final stochastic component, ε, must likewise not correlate with regressors, but we will assume so henceforth. Note also that lack of correlation of u_{0j} and u_{1j} with regressors is assumed in all the estimation strategies discussed here.

3.3 *Separate-Subsample vs. Dummy-Interaction Estimation*

Jusko and Shively (in this issue) offer several arguments that may favor a two-step strategy. First, intrinsic interest in distinct models for each macro-level context dictates that researchers actually do want estimates of β_{0j} and β_{1j} per se as well as merely en route to estimates of micro-macro interactive effects, γ_{11} (and possibly macro-level coefficients, γ_{01}, also). Regressing y_{ij} on x_{ij} separately in each of the j subsamples will produce such estimates, but so too would regressing all y_{ij} on the complete set of j macro-level indicators and interactions of each of those with x_{ij}. Indeed, as is well known, either procedure, *separate-subsample* or call it *dummy-interaction*, produces mathematically *identical* estimates of β_{0j} and β_{1j}. Thus they share the same bias, consistency, and efficiency properties and so would serve equally well (for their part: standard errors differ as seen below) in any subsequent analysis relating them to contextual factors, z_j. Likewise, either procedure equally easily accommodates macro-unit-specific regressors such as, say, respondent ethnic-group indicators for ethnicities that do not exist in all j.[14] In terms of these β_{0j} and β_{1j} coefficient estimates alone,[15] then, whether to dummy-interact and pool or estimate in separate subsamples is wholly irrelevant or purely a matter of practical implementation ease.[16]

Standard errors, however, will differ by these procedures. First, notice from Eq. (5) that either option, by allowing β_{0j} and β_{1j} to vary arbitrarily across j, assures that the macro-unit-specific shocks will be identically zero: $u_{0j} = u_{1j} = 0$. Accordingly, the sole stochastic term in either case is ε_{ij}. To this, pooled OLS applies $\hat{V}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$, with $s^2 = \frac{1}{n-k}\sum e_i^2$, across the full sample, whereas separate-subsample OLS applies $\hat{V}(\hat{\boldsymbol{\beta}}_j) = s_j^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$, with $s_j^2 = \frac{1}{n_j - k_j}\sum e_{i\in j}^2$, separately in each j. Given the block diagonality of $(\mathbf{X}'\mathbf{X})^{-1}$ in the dummy-interaction model, the jth block of $(\mathbf{X}'\mathbf{X})^{-1}$ in $\hat{V}(\hat{\boldsymbol{\beta}})$ exactly equals the $(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ in $\hat{V}(\hat{\boldsymbol{\beta}}_j)$, so only the s^2 vs. s_j^2 differ in the standard-error estimates. The former assumes constant variance and no correlation across all j, but the latter only within each j, leaving unspecified any cross-subsample heteroskedasticity or correlation. Because $\hat{\boldsymbol{\beta}}$ and $\mathbf{X}$ and $\mathbf{y}$ are identical in either procedure, the e_{ij} are also identical. Given that, and with $n = \sum n_j$ and $k = \sum k_j$, s^2 is the average across j of s_j^2 (i.e., $s^2 = \bar{s}_j^2$). If the residuals are truly homoskedastic across contexts j, then the s^2 from pooling and so $\hat{V}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$ is *efficiently* constant and s_j^2 and $\hat{V}(\hat{\boldsymbol{\beta}}_j) = s_j^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ *inefficiently* vary across contexts. Recapturing this efficiency by constraining $s_j^2 = \bar{s}_j^2$ in separate-subsample estimation would be extremely difficult, likely requiring some recursive estimation strategy. Unequal variance across j is perhaps more plausible, though. If so, then s_j^2 and $\hat{V}(\hat{\boldsymbol{\beta}}_j) = s_j^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ from the unit-by-unit *OLS* are *correctly* variant, and s^2 and $\hat{V}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$ from pooling *incorrectly* constant, i.e., biased and inconsistent (although right on average across j). Once again, redressing this bias/inconsistency requires only simple application of robust standard-errors (plain heteroskedasticity-consistent Eq. [12] should suffice here) and/or FGLS (j indicators suffice for regressors in the auxiliary regression).

---

[14]For example, to include a Basque indicator in Spain only, simply enter it only in that regression by the two-step procedure and include the Basque indicator, which is strictly zero outside Spain, in the whole sample for the pooled procedure.

[15]Whether in one-step pools or two-step separate subsamples, what controls to apply and, more generally, ensuring that β_{0j} and β_{1j} estimate the same substantive/theoretical quantities across all j, is paramount. However, it is equally so in any strategy, and no strategy has any means of ensuring this theoretical issue empirically, so this issue cannot serve to evaluate alternative estimation strategies.

[16]For example, some software facilitates indicator and indicator-interaction generation; others facilitate sample restrictions on repetitions of the same estimation command.

Finally, $V(\varepsilon)$ may have nonzero off-block-diagonal elements, meaning residuals exhibit some cross-subsample correlation. In our substantive examples, time periods of inexplicably large/small deficits in some countries may correlate with deficits elsewhere, or inexplicably warm/cold feelings toward right parties in some countries may correlate with feelings toward others elsewhere. If so, either separate-subsample or pooled OLS will produce inefficient coefficient estimates and biased, inconsistent, and inefficient standard-error estimates. Intuitively: some cross-subsample information exists that either OLS procedure ignores. Again, attempting to incorporate such cross-subsample information (correlation) to enhance efficiency or adjust standard errors would require some difficult recursive-iteration strategy in true separate-subsample estimation, although a related strategy like seemingly unrelated regression (SUR) might suffice. In pooled samples, incorporating correlation information to enhance efficiency and improve standard-error estimation is just another application of FGLS (e.g., Parks procedure if degrees of freedom suffice, or some more-limited parameterization thereof if not) and/or consistent standard errors. Beck and Katz (1995, 1996) panel-corrected standard errors (PCSE) would suffice to render standard errors consistent to one common form of cross-subsample correlation (contemporaneous correlation in a TSCS context).[17]

Thus, for estimating β_{0j} and β_{1j} *per se* and their standard errors given some intrinsic interest in such context-unique models of politics, little distinguishes pooled dummy-interaction from separate-subsample strategies. Just the assumptions about $V(\varepsilon)$ might be distinct. (In models with stochastic components fully determined by mean parameters, like binary or count models, not even this differs. There, likelihood maximization yielding identical $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ by either strategy implies that it also produces identical standard-error estimates.[18]) Pooled OLS does impose likely implausible variance restrictions, but simple FGLS or robust standard-error estimators will adequately redress this and, indeed, can be specified to reproduce separate-subsample assumptions, and so estimates, exactly. Cross-subsample information, on the other hand, is inherently difficult for two-step strategies to accommodate, whereas it is simpler (some standard FGLS or robust procedure will usually apply) in pooled-sample estimation. Therefore, absent cross-subsample information, two-step strategies might be at best marginally easier, not requiring FGLS weighting or robust standard errors, whereas with such information, pooled estimation is unambiguously much easier. However, for purposes of intrinsic interest in obtaining j unique models, the choice is mostly (in models where means and variances are codetermined, like binary or count: *purely*) one of personal taste or software facilities.

For purposes of estimating the effects on the outcome, y_{ij}, of micro- and macro-level factors, x_{ij} and z_j, and their interaction, $x_{ij}z_j$, however, either separate-subsample and pooled-dummy-interaction estimations are mere preliminaries. The researcher seeks estimates of Eqs. (6)–(8), i.e., the effects on outcomes like fiscal policies or party affinity of micro-level factors like partisanship or income, $\frac{\partial y}{\partial x}$, of macro-level factors like district magnitude or inequality, $\frac{\partial y}{\partial z}$, and if and how effects of these micro- and macro-level factors depend on the other variable, $\frac{\partial^2 y}{\partial x \partial z} \equiv \frac{\partial^2 y}{\partial z \partial x}$. Neither of these strategies yields direct estimates of any of the parameters of interest (γ_{00}, γ_{01}, γ_{10}, and γ_{11}), though. These are obtained instead in second-stage estimations wherein the $\hat{\beta}_j$ from the first stage are regressed on z_j as shown in the Jusko and Shively (2005) and Lewis and Linzer (2005) contributions to this

---

[17]In *Stata*, one uses the *xtpcse* command to obtain these.

[18]Intuitively: the same coefficient estimates emerge from the same (parts of) likelihood functions being maximized at the same points. Since standard errors are curvatures of those likelihood functions at those points, they are also identical.

issue and applied in most others. For these more theoretical purposes, except for the minor differences discussed above, full-dummy-interaction and separate-subsample estimation strategies are identical; they each provide estimates, equally good ones usually, of $\hat{\beta}_j$ and $\hat{V}(\hat{\beta}_j)$ to a second stage estimating the actually desired $\hat{\gamma}$.

3.4   *One-Step vs. Two-Step Estimation*

What are the trade-offs, then, between two-step strategy,

$$y_{ij} = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_{ij} + \hat{\varepsilon}_{ij}; \quad \hat{\beta}_{0j} = \hat{\gamma}_{00} + \hat{\gamma}_{01}z_j + \hat{u}_{0j}; \quad \hat{\beta}_{1j} = \hat{\gamma}_{10} + \hat{\gamma}_{11}z_j + \hat{u}_{1j}, \quad (16)$$

and one-step estimation of the pooled linear-interactive model (14). First, realistically, let the linear interaction $x_{ij}z_j$ only partially capture macro-level variation in the effect of x_{ij}, and z_j only partially capture macro-level variation in the conditional mean, and simplify by assuming ε_{ij} uncorrelated with regressors and $V(\varepsilon)$ spherical. We have already shown pooled estimation of Eq. (14) unbiased and consistent for $\hat{\gamma}$ if $E(u_{0j}, u_{1j} \mid x_{ij}, z_j) = 0$, but that inefficiency and incorrect standard errors arise even if $V(u_{0j})$ and $V(u_{1j})$ are spherical. We also saw simple FGLS and consistent-standard-error strategies for redressing these shortcomings. Jusko and Shively (2005) show, conversely, that two-step estimation of Eq. (16) can produce unbiased and consistent coefficient estimates and standard errors without efficiency costs relative to pooled estimation under broad circumstances, the most important of which is the absence of cross-subsample information as mentioned above and elaborated below.

One aspect of the trades between the one- and two-step approaches is that, of the three effects of interest to multilevel modelers, only the micro-level effects, $E(\frac{\partial y_{ij}}{\partial x_{ij}}) = \gamma_{10} + \gamma_{11}z_j$, and the dependence of the micro- and macro-level effects on each other, $E(\frac{\partial^2 y_{ij}}{\partial x_{ij}\partial z_j}) = \gamma_{11}$, emerge directly from a single-equation estimation in the second to the two-stage steps. Those two parameter estimates emerge by regressing $\hat{\beta}_{1j}$ on (a constant and) z_j, weighted by the estimated variance of $\hat{\beta}_{1j}$ (Jusko and Shively 2005; Lewis and Linzer 2005). To obtain macro-level effects, $E(\frac{\partial y_{ij}}{\partial z_j}) = \gamma_{01} + \gamma_{11}x_{ij}$, however, one must estimate the last *two* expressions in Eq. (16). Furthermore, because $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ necessarily correlate, one must regress $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ on (constants and) z_j in a *simultaneous* system of equations, weighted by the estimated variance-*covariance* of those two coefficients. The one-step estimation, contrarily, produces all three parameters of interest (plus the conditional mean) directly in its one stage.

Another aspect of the trades, however, is that, if the contextual factor(s), z_j, leave some macro-unit-specific variation unaccounted, the one-step estimator must rely on the orthogonality of those unaccounted u_{0j} and u_{1j} for the unbiasedness and consistency not only of the $\hat{\gamma}$ estimates, but also for the unbiasedness of coefficients on other explanators. The two-step estimator also relies for the unbiasedness and consistency of its $\hat{\gamma}$ estimates on this orthogonality; without that, the $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ estimates from the first step will contain stochastic components related to z_j, which will bias the second-stage estimates of $\hat{\gamma}$ exactly as in the one-step estimation. However, because the arbitrary cross-context variation $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ will absorb any macro-unit-specific variation, including that unexplained by z_j, the first step of the two-step estimator produces estimates of other coefficients in that first stage that will not be biased by the failure of z_j to account all cross-contextual variation.[19]

---

[19]This is essentially the standard argument for fixed effects, sharing its strengths and weaknesses.

A third aspect of the trade-offs is the *possibly* relatively heavier reliance of one-step estimators on large samples for accurate standard-error estimation (and efficiency) when z_j only partially models cross-contextual variation. FGLS, e.g., relies on estimating few variance-covariance parameters relative to degrees of freedom for its optimum efficiency and standard-error-estimation properties. Consistent variance-covariance estimators rely on asymptotics by definition. Pure heteroskedasticity-consistent robust standard errors, Eq. (12), such as those appropriate in nonclustered random-effects conditions or in the full-dummy-interaction case, seem to function reasonably well even in fairly small samples. Their performance, as crudely gauged by suggested small-sample corrections to the base estimator, should be a decreasing function of $N/(N - k)$, such that even $N = 55$, $k = 5$ would yield only about 10% overconfidence. Existing simulations support this intuition.

Clustered-heteroskedasticity-consistent standard errors, however, have asymptotics in both i and j dimensions, their performance by such gauges being a decreasing function of $[J/(J - 1)][N/(N - k)]$. As this would suggest, $J = 50$, $n_j = 100$ seems comfortable in a linear regression model, judging by Franzese and Kam's (2005) simulations. Eduardo Leoni (2005) explored clustered standard errors for logit, however, finding hypothesis-test rejection rates about 2.5 times too high (.13 for .05) for $J = 24$ and coverage rates suggesting 17% overconfidence. Small-sample adjustments for maximum likelihood are $[J/(J - 1)]$, which suggests only about 5% overconfidence, so clustered-standard-error strategies may lean even more heavily on large J in contexts beyond linear regression. However, Leoni's simulations also showed that a robust-cluster estimator applying small-sample corrections performed remarkably well. Coverage rates for that adjusted clustered-standard-error estimator were 9%, 5%, 1%, and 3% *too large* in samples of $J = 10, 15, 20$, and 40 and just 1% and 2% too small in samples of $J = 30$ and 500. We need more simulations to understand the small-sample properties of the consistent-standard-error estimates often needed for single-stage estimation of multilevel models, and of which small-sample corrections work best under what conditions, but these results are very encouraging. Two-stage estimators, of course, also rely on large samples for small standard errors, but standard-error accuracy *may* lean less heavily on large macro-unit samples, depending on how few parameters the FGLS (or consistent-standard-error) estimation(s) in their second steps require.[20]

Penultimately, and most favorably for two-step approaches, are the implications of incomplete contextual-variation modeling by z_j in cases in which stochastic and systematic components are not additively separable. If, for example, the multilevel model is of the following probit form,

$$y_{ij} = \Phi(\beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}); \quad \beta_{0j} = \gamma_{00} + \gamma_{01}z_j + u_{0j}; \quad \beta_{1j} = \gamma_{10} + \gamma_{11}z_j + u_{1j}, \quad (17)$$

then we cannot separate the estimation of the systematic coefficients of x_{ij}, z_j, and their interaction from the estimation of the standard errors of those coefficients. Accordingly, the relatively benign effects of context-specific error components seen in the linear Eqs. (10)–(11), being confined to largely rectifiable coefficient inefficiency and standard-error inaccuracy, may not generally hold in this case. Confidence that z_j adequately models the fuller extent of cross-contextual heterogeneity must be higher. Lacking such confidence, the appropriate one-step recourse would be to write the full likelihood function Eq. (17)

---

[20]One-step strategies only *possibly* lean more, and two-step *may* lean less, heavily on large samples, but this has not been demonstrated. Since, in principle, equivalent one- and two-step strategies exist for any empirical task and data properties, I rather suspect that, in fact, any differences are illusory.

implies, which would mean nesting two normal p.d.f.'s (assuming that the second and third expressions are independent linear-normal equations) within a binomial p.d.f. (probability density function), which, while feasible, could require programming and yield an unpleasant likelihood surface to search. Two-step estimation, contrarily, is greatly aided in this regard by the asymptotic normality of maximum-likelihood estimates. The $\hat{\boldsymbol{\beta}}$ estimates from its first step probits that serve as dependent variables in the second step will be (asymptotically) normally distributed, and so ideal for linear regression, just by virtue of having emerged from the first-stage maximum-likelihood estimation.[21]

Finally, and least favorably for two-step estimation, cross-subsample information, as indicated already, is usually difficult for two-step estimators to incorporate but simple for one-step estimators. If, for example, we know that some coefficients are equal across macro-level subsamples or equal across micro-level units, or if we wish to constrain them to be so to enhance efficiency believing it "not far from true," we can do so in one-step estimation simply by including that variable in the model and not interacting it with x_{ij} or z_j, respectively. In separate-subsample estimation, allowing some macro-level factors to have equal effect on all micro units ij is almost as easy; one simply includes these z_j terms only in the $\hat{\beta}_{0j} = \hat{\gamma}_{00} + \hat{\gamma}_{01}z_j + \hat{u}_{0j}$ and not the $\hat{\beta}_{1j} = \hat{\gamma}_{10} + \hat{\gamma}_{11}z_j + \hat{u}_{1j}$ second-stage regression. Constraining some micro-level factors to have equal, proportional, or otherwise related coefficients across contexts j, though, is far harder. As with cross-context residual variance or correlation discussed above, two-step estimation requires some iterated strategy to accommodate such cross-sample information. Furthermore, some sorts of cross-sample information, such as systematic interdependence across j in the outcomes y_{ij}, which our cross-national fiscal-policy example would likely exhibit, e.g., will not only render separate-subsample standard-error estimates biased and coefficients inefficient but also will bias coefficient estimates. As Franzese and Hays (2005) show regarding interdependence in economic policy making, for example, estimating, say, French or Ohioan fiscal-policy models ignoring the feedback from, say, German and Michiganian fiscal policy will produced biased models of French or Ohioan fiscal policy.[22]

To summarize the comparison of separate- and pooled-sample estimation strategies, then:

(a) In principle, what one can do in separate subsamples in two steps one can also do in one step with interactions (etc.), and *vice versa*, but some things are easier one way or the other.

(b) Separate-subsample and full dummy interactions produce identical coefficient estimates and standard errors that, at most, rectifiably differ; with nonseparable error components, the two are exactly identical. Either can handle context-specific regressors equally easily.

(c) Separate-subsample coefficient estimates are at least inefficient (biased also in some cases) and standard errors inaccurate relative to appropriate pooled estimators if cross-subsample information exists: homoskedasticity ($V(\{\varepsilon_{ij}\}) = \sigma^2 \boldsymbol{I}$), correlation

---

[21]Again, they only "*could* be" ugly and unpleasant because these are tastes and, anyway, have not been shown, and, again, I suspect otherwise (see note 20). Because one- and two-step strategies are essentially alternative orderings of the same tasks, I suspect these likelihoods will, in fact, be well behaved and easily searchable by the same logic that proves ML estimates asymptotically normal.

[22]The bias here is textbook omitted-variable bias: German policy (inter alia) causes French policy and likely correlates with other causes of French policy; therefore, estimating French policy models ignoring German ones is biased. Correctly estimating such interdependence models is greatly challenging in its own right, but the point here is simply that ignoring this sort of cross-sample information creates greater problems than mere inefficiency and fixable standard-error inaccuracies.

($V(\{\varepsilon_{ij}\})$ not block diagonal), or coefficients relate across j. Leveraging such cross-equation information is usually difficult in two-step estimation.

(d) Two-step estimation of macro-level effects requires second-stage systems of equations.

(e) Two-step estimation yields coefficient estimates for micro-level variables that are more robust to misspecification of macro-level effects than one-step linear-interaction models.

(f) One-step linear-interaction models *may* rely more heavily on large samples, especially in macro-unit dimensions, for standard-error accuracy and efficiency enhancement.

(g) One-step linear-interaction models struggle to incorporate the stochastic interactions that are plausible in multilevel models into outcomes whose systematic and stochastic components are nonseparable. Two-step estimation actually facilitates this.

4 Comparative Time-Series Cross Sections, Comparative Panels, and Comparative Surveys

In conclusion, consider, in light of the preceding discussion, the typical properties of panel and TSCS data in comparative and international politics and political economy and those of panel and cross-context survey compilations in comparative microbehavioral research. The key issues, as noted at the outset and seen throughout the article are the dimensions, the likely systematic and stochastic properties, and the research questions in the sample and substantive area.

In comparative micro-behavioral research, such as exemplified throughout this issue, datasets are commonly large (hundreds to thousands of observations), independently randomized surveys of individuals pooled across a few countries (10 to 20 or, more rarely, 30) or subnational political units (perhaps 50). Being so large within each macro level, i.e., having such large I (plus independent i), efficiency at micro levels is unlikely to be of much concern. Moreover, because the surveys are being *independently* randomized, one expects quite limited, if any, cross-sample information. At most, individual opinions (e.g., party affinities) in France, *in their aggregate*, might affect individual opinions in Germany. In principle, any estimator that ignores such cross-subsample information, if it exists, will suffer bias and inconsistency, as well as inefficiency (Franzese and Hays 2005). In this context, however, such information is unlikely to be sizeable[23] if it exists at all. Moreover, this (small-to-zero) dependence of each respondent in one j on the *aggregate* of respondents in each of the other j would be absorbed in the macro-level-specific constants, which, while problematic for second-stage estimates, should not therefore hamper first-stage estimates. Furthermore, micro-behavioral research likely stresses micro-level effects, $\frac{\partial y_{ij}}{\partial x_{ij}}$, and their context-conditioning, $\frac{\partial^2 y_{ij}}{\partial x_{ij} \partial z_j}$, much more than macro-level effects, $\frac{\partial y_{ij}}{\partial z_j}$. Finally, outcomes are far more often qualitative than linear-continuous. Thus, referring back to our summary comparison of separate-sample versus pooled-interactive strategies, we see that, along every consideration, typical conditions in cross-context pooled and independently randomized survey analysis in comparative micro-behavioral research are ideal for

---

[23]Each element of each off-diagonal block in the overall variance-covariance matrix would reflect the covariance of *one* (random) respondent in the row block with *one* (random) respondent of the column block and so would be some very small number (the same very small number for each element in that block).

two-step strategies (with the first step being equally well served by full-dummy-interaction or separate-subsample estimation).[24] Insofar as Leoni's simulations apply, pooled-interaction with clustered-heteroskedasticity strategies could also work in these conditions if small-sample adjustments are applied.[25]

In typical survey panel data, in which large numbers (again, hundreds to thousands) of the same respondents are observed small numbers (usually fewer than ten, very rarely perhaps twenty or thirty) of times, conversely, all the considerations weigh oppositely. The very small number of micro-level observations will often render separate subsample estimation literally impossible (negative degrees of freedom), and, even where possible, capitalizing upon cross-subsample information will be indispensable to reasonable efficiency. Cross-subsample (here, cross-respondent) information is hopefully sizeable because panel-data analysts must lean on it heavily. Notice, however, that here cross-subsample information is, substantively, precisely the same cross-respondent information that is all that exists at the micro-level in randomized survey data. That is, assuming an effect is constant across subsamples in the panel-survey-data case is the same as estimating that one effect at all in nonpanel survey data. Furthermore, the very large number of macro-level units implies that consistent standard-error strategies (or FGLS in linear regression) will work quite well.[26]

Finally, in the TSCS data typical of comparative and international politics and political economy, observations are usually of political (national or subnational) units at the macro level, over time at the micro level. Both micro and macro levels tend to have intermediate numbers of observations. In comparative/international political economy, developed or developing country samples typically have 15–35 macro-level units; global samples might triple or quadruple this; micro-level units (time, usually years) typically vary from 10 to 40. International relations contexts might have global samples of countries, dyads, or directed dyads and widely varying years (from 10 to over 100); sometimes macro levels are confined to *relevant* or *great-power* cases, leaving J as few or fewer countries or (directed) dyads in the 15–30 range or lower. Therefore, leveraging cross-subsample information in such contexts, as with panel surveys, is at least extremely useful, usually crucial, and occasionally indispensable. As a matter of substance as well as necessity, moreover, observations are highly likely to be related across macro-level contexts. Indeed, in many political economy and international relations contexts (e.g., globalization and strategic relations), cross-unit *interdependence* is substantively central. Finally, macro-level effects, $\frac{\partial y_{ij}}{\partial z_j}$, are usually at least as central as micro-level effects, $\frac{\partial y_{ij}}{\partial x_{ij}}$, and their context conditioning, $\frac{\partial^2 y_{ij}}{\partial x_{ij} \partial z_j}$, in these research areas. In such conditions, one-step estimation strategies seem the better option. They may face some challenges in accurate standard-error estimation, but these seem surmountable with *small-sample-adjusted* consistent-estimator or FGLS strategies, especially since dependent variables are more often linear continuous. Unbiased coefficient estimation seems on stronger ground for the same reason.

---

[24]Duch and Stevenson (2005) is a partial exception. Surveys across elections within countries are three levels. Cross-election-within-country information (intermediate level) is likely much greater, recommending extra efforts in this direction.

[25]Shrinkage estimators like HLM, conversely, are unlikely to shrink much from the *within* estimates given very large I and much smaller J, so they would tend to serve little practical purpose here (Beck and Katz 2005; see the appendix to this article on the *Political Analysis* Web site).

[26]Here, shrinkage estimators likely shrink separate-subsample estimates almost fully to the between estimator, since J is large and I so small; again, shrinkage estimators may serve little practical purpose except where the I (time) dimension extends sufficiently to allow reasonable subsample estimates (see the appendix on the *Political Analysis* Web site).

Two-step procedures, on the other hand, are unlikely to prove effective or practical (or even, in some cases, possible).[27]

References

Beck, N., and J. Katz. 1995. "What To Do (and Not to Do) with Time-Series-Cross-Section Data in Comparative Politics." *American Political Science Review* 89(3):634–647.

Beck, N., and J. Katz. 1996. "Nuisance or Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6:1–36.

Beck, N., and J. Katz. 2005. "Random Coefficient Models for Time-Series-Cross-Section Data." Presented at the 2001 meetings of the Political Methodology Organization Section of the American Political Science Association.

Bowers, J., and K. Drake. 2005. "EDA for HLM: Visualization When Probabilistic Inference Fails." *Political Analysis* doi:10.1093/pan/mpi031.

Brambor, T., W. R. Clark, and M. Golder. 2005. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* doi:10.1093/pan/mpi014.

Davidson, R., and J. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Franzese, R., and C. Kam. 2005. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis: A Refresher and Some Practical Advice*. Unpublished manuscript. (Available at www.personal.umich.edu/~franzese/Interactions_Michigan.030305.pdf.)

Franzese, R., and J. Hays. 2005. *Spatial Econometric Models for Political Science*. (Available at www.personal.umich.edu/~franzese/FranzeseHays.SpatialEcon.Book.pdf.)

Greene, W. H. 2003. *Econometric Analysis*. Upper Saddle River, NJ: Pearson Education, Inc.

Jusko, K. L., and P. Shively. 2005. "A Two-Step Strategy for the Analysis of Cross-National Public Opinion Data." *Political Analysis* doi:10.1093/pan/mpi030.

Leoni, E. 2005. "How to Analyze Multi-Country Survey Data: Results from Monte Carlo Experiments." Paper presented at the 2005 Midwest Political Science Association Conference.

Lewis, J., and D. Linzer. 2005. "Estimating Regression Models in which the Dependent Variable Is Based on Estimates." *Political Analysis* doi:10.1093/pan/mpi026.

---

[27]Here, finally, shrinkage estimators may serve a practical purpose more often as, with both J and I intermediately sized, estimates may differ meaningfully from both within- and between-estimator extremes (see the appendix on the *Political Analysis* Web site).