Advance Access publication March 7, 2007

Political Analysis (2007) 15:165–181
doi:10.1093/pan/mpm006

A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data

Boris Shor
Harris School of Public Policy Studies, University of Chicago,
1155 E. 60th Street, Suite 185, Chicago, IL 60637
e-mail: bshor@uchicago.edu (corresponding author)

Joseph Bafumi
Department of Government, Dartmouth College,
6108 Silsby HallHanover, NH 03755
e-mail: joseph.bafumi@dartmouth.edu

Luke Keele
Department of Political Science, Ohio State University,
2137 Derby Hall, 154 N Oval Mall, Columbus, OH 43210
e-mail: keele.4@polisci.osu.edu

David Park
Department of Political Science, George Washington University,
1922 F Street, N.W. 414C, Washington, DC 20052
e-mail: dkp@gwu.edu

The analysis of time-series cross-sectional (TSCS) data has become increasingly popular in political science. Meanwhile, political scientists are also becoming more interested in the use of multilevel models (MLM). However, little work exists to understand the benefits of multilevel modeling when applied to TSCS data. We employ Monte Carlo simulations to benchmark the performance of a Bayesian multilevel model for TSCS data. We find that the MLM performs as well or better than other common estimators for such data. Most importantly, the MLM is more general and offers researchers additional advantages.

1 Introduction

Time-series cross-sectional (TSCS) data have become increasingly popular over the last 10 years in political science. With TSCS data, researchers can test theories about both cross-sectional and cross-temporal variation. But this flexibility comes at the cost of methodological complexity. TSCS data structures invariably violate the standard assumptions

Authors' note: A previous version of this article was presented at the 2005 Midwest Political Science Meeting. We would like to thank the following for comments and advice in writing this paper: Andrew Gelman, Nathaniel Beck, Greg Wawro, Sam Cooke, John Londregan, David Brandt. Any errors are our own.

underlying classical linear regression models. Much effort has been made to develop models that can handle the pitfalls of such data.

Analysts with TSCS data are faced with a variety of estimators. The most popular options for analyzing such data are variations on ordinary least squares (OLS). In particular, analysts tend to use either fixed effects (FE) or "panel-corrected standard errors" (PCSEs) (Beck and Katz 1995; Beck 2001) with OLS. These alternatives have significantly improved upon prior estimators like feasible generalized least squares, especially in the small data sets likely to be employed by political scientists.

These advances in modeling TSCS data have come from a recognition that merely pooling observations across time in addition to space is problematic. We show that thinking about such data in terms of a MLM is an even more fruitful perspective. In other words, these data have an explicit structure that should be modeled. The advantages of such an approach include vastly increased flexibility in model specification, potential improvements in model fit, and better accounting of uncertainty at all levels of analysis.

MLMs have seen increased use recently in political science, yet they have rarely been applied to TSCS data.[1] Multilevel modeling offers substantial advantages over existing methods for such data, as we will show. A MLM can be estimated with a variety of well-understood techniques, including variations on maximum likelihood (ML) and Bayesian inference. We choose to explore the use of a Bayesian multilevel model (BML) because it offers a number of advantages, primarily in providing analysts with flexibility for modeling complex error structures and contextual data that are characteristic of TSCS data. Yet, we must emphasize that maximum likelihood MLMs also offer big advantages over existing techniques. In what follows, we discuss past work using MLMs with TSCS data, outline the structure of a BML, and use Monte Carlo experiments to explore its properties relative to other popular estimators for such data.

1.1 *Multilevel Models*

MLMs are a generalization of standard regression techniques, and as such can generically be expected to perform as well, or better (Gelman 2006; Gelman and Hill 2006). A multilevel structure to modeling TSCS data is natural, given the nonindependence of observations across time and space and the prevalence of unit heterogeneity. Yet such work is a rarity in the literature.

One major exception is Western (1998), where he fits a Bayesian hierarchical[2] model to predict economic growth in Organisation for Economic Cooperation and Development countries. Western (1998) shows that his hierarchical model provides the following: (a) more accurate forecasts than other models, (b) more accurate estimates of time-series effects than an unpooled analysis, and (c) more realistic accounting of uncertainty than a conventional pooled analysis. However, the use of a single data set does not allow us to draw general conclusions about the appropriateness of a MLM for such data.[3]

Similarly, Shor et al. (2003) test the properties of a BML on U.S. federal spending data across states and years. They find that this model performs well relative to conventional estimators for this particular empirical example. However, they do not provide a systematic exploration of the model's properties.

[1] An exception is Western (1998), but he does so only in the context of a single data set.

[2] Hierarchical models are a type of MLM where observations are completely and uniquely nested within higher-level units.

[3] In addition, Western (1998) does not include an indicator for time. Failure to include a time indicator leads to missing the common effects of particular years across all cross-sectional units for that year. These time shocks are a source of contemporaneous correlation. We include these time indicators in our MLM.

Beck and Katz (forthcoming) use simulations to find that ML estimators for random coefficient models are superior in terms of efficiency to various types of pooled and unpooled OLS estimators in small samples. However, they investigate a limited set of data generating processes (DGP).

What is missing is a study of the performance of MLMs when important components of the TSCS DGP and error structure vary. Using a set of Monte Carlo experiments, we seek to understand how well a MLM performs given common challenges with such data. We benchmark the results against other widely used estimators for such data. This approach allows us to acquire more confidence about the general applicability of multilevel modeling applied to these data.

MLMs—which incorporate data at multiple levels of analysis simultaneously—have been introduced and applied in political science research (Gelman and King 1993; Western 1998; Park 2001; Steenbergen and Jones 2002; Gelman et al. 2003; Shor et al. 2003; Shor 2004, 2006; Bafumi 2004a, 2004b; Park, Gelman, and Bafumi 2004; Bafumi et al. 2005; Gelman and Hill 2006; Gelman et al. forthcoming).[4] For example, states and years can be modeled to predict presidential electoral outcomes (Gelman and King 1993), individual- and state-level covariates can be simultaneously modeled to predict issue attitudes by state (Park, Gelman, and Bafumi 2004), congressional district, state, and longitudinal predictors can be used to predict geographic patterns of federal spending (Shor 2004), or individuals, nations, and parties can be modeled to predict support for European integration (Steenbergen and Jones 2002). This approach was popularized in educational research, where it was used to explain the variation in student performance using student-level information as well as school-level variables (Raudenbush and Bryk 2002).

To demonstrate how multilevel modeling works compared to commonly applied alternatives, let us consider three procedures for estimating group effects with TSCS data, where the cross-sections (units) are American states. In the equations below, let y_t be the response variable, x_t an independent variable, β a parameter of interest to be estimated, γ_t an indicator for each time period, α an intercept parameter, and ε_t an error term.

Procedure 1. Ignore states (complete pooling)

$$y_t = \alpha + \gamma_t + \beta x_t + \varepsilon_t. \tag{1}$$

Procedure 2. Include states but ignore the state-level error (no pooling)

$$y_{jt} = \alpha_j + \gamma_t + \beta x_t + \varepsilon_{jt}. \tag{2}$$

Procedure 3. Include states and add state-level error (partial pooling)

$$y_{jt} = \alpha_j + \gamma_t + \beta x_t + \varepsilon_{jt}, \tag{3}$$

$$\alpha_j = \alpha_0 + \eta_j, \tag{4}$$

$$\text{where } \eta_j \sim N(0, \sigma_\eta^2).$$

The term of interest in the above equations is α. In the first, states are pooled together (equation 1, or "complete pooling"). Under complete pooling, α is assumed to be constant

[4]Also see the Autumn 2005 edition of *Political Analysis* devoted to the analysis of multilevel data sets.

across all the group-level units, which in this example are states. That is, it is assumed that there is no variability at the state level. If this assumption is incorrect, as it often is in practice, estimates of the standatd errors (SEs) for the model will be flawed.

At the opposite extreme, we may fit a separate regression model within each state (equation 2, or "no pooling"). Under no pooling, α is now subscripted by j since we estimate a separate intercept for each state. State-level indicators are now estimated with maximal variance, which will often be unreflective of population parameters. This is sometimes called a FE specification. It does not allow the sharing of potentially important information from across states.

We now turn to equation 3, or "partial pooling."[5] Under partial pooling, α_j is now an outcome in the model. The term α_0 represents the average intercept across states, whereas η_j is the unique effect of state j on α that is assumed to be a random shock from a normal distribution.

MLMs allow the variance of the unit effects, α_j, to be estimated conditional on the data and parameters at *all* levels. By taking advantage of the error structure at different levels of the equation, a within-unit and between-unit variance is estimated. In this process, the units are allowed to "shrink" or gravitate to a common mean that in turn often produces more accurate estimates of these effects. For example, if the units are states and one state has little data with severe outliers, the estimate of the effect for this state is shrunk to the common mean of all states. The amount of shrinkage is inversely proportional to the amount of data available; more data means less movement toward the overall mean.

When analyzing a sample with commonly distributed units, partial pooling offers potentially great advantages since incomplete knowledge of the population parameters can be at least partly compensated by borrowing information across units. Since the variance across units is estimated, MLMs should do no worse than applying a complete pooling approach to data with no unit effects, or an FE approach where the units in the data are highly variable. In between these two extremes, it should do better (Raudenbush and Bryk 2002; Gelman et al. 2003).

If, however, those units do not come from a common distribution (e.g., pooling together U.S. states and Canadian provinces), then a practitioner may achieve more accurate estimates without partial pooling. In order for the estimates of unit effects to benefit from partial pooling, the assumption that they come from a common distribution must hold. On the other hand, the alternatives of complete pooling are worse under this scenario, and no pooling may prove to subset the data to levels where reasonable estimates are impossible (Bartels 1996). MLMs offer a way around this dilemma, by allowing practitioners to assign different prior distributions to units in accordance with different a priori expectations.

In TSCS data, collinearity is frequently present (Western and Jackman 1994; Western 1998). Political scientists employing TSCS data are often interested in slow moving or completely time-invariant predictors that vary only cross-sectionally. Examples include the presence of democracy in comparative politics data, or a particular political institution specific to a state in American politics data. Yet, because of collinearity, the inclusion of these predictors is problematic in the presence of FE. The variability of the effects will often be exaggerated, yielding incorrect parameter estimates for the second-level predictors. This is not a problem in the multilevel framework, since partial pooling allows estimates of units to borrow strength from the whole sample and shrink toward a common mean.

MLMs generally provide improved model fit relative to least-squares–based estimators. When the number of observations within units is low,[6] the partial pooling design that

[5]An earlier exploration of partial pooling can be found in Bartels (1996).

[6]In the case of unbalanced data, even if the average number of observations is acceptable, individual units may have too few observations.

multilevel modeling makes possible provides more reasonable parameter estimates than pooled or unpooled designs. This is because complete pooling ignores important local variation, whereas no pooling is subject to the effect of outliers (Gelman and Hill 2006). Partial pooling offers a middle ground where each estimated parameter has the potential to borrow strength from other parameters in the model. This is important because least-squares estimators are justified by asymptotics in time. In sum, we should expect that MLMs do better for small-*T* (or -*N*) data sets than OLS-like estimators (Gelman et al. 2003; Gelman and Hill 2006).

1.2 *Bayesian Inference*

In the past decade or so, the use of Bayesian inference has advanced rapidly in the social sciences.[7] Essentially, the Bayesian framework combines prior (nonsample) information with sample data to produce a posterior distribution of parameter estimates. The mean and standard deviation (SD) of this distribution are then comparable to OLS and ML parameter estimates. We rely on the Gibbs sampler to fit the Bayesian TSCS model. The Gibbs sampler partitions the set of unknown parameters and then estimates them one at a time, or one group at a time with each parameter or group of parameters estimated conditional on all the others (Gelman 2006). The algorithm allows us to easily estimate separate parts of the model, even if it is difficult to estimate all the parameters at once.

MLMs may be estimated using ML (Raudenbush and Bryk 2002) or Bayesian simulation (Gelman et al. 2003). We know that ML estimates have desirable properties: they are consistent, unbiased, and efficient. Combined with advances in statistical packages like *lmer()* in *R* and *xtmixed* and *gllamm* in *Stata*, it is natural to wonder why we turn to Bayesian inference to estimate these models. We provide some of those reasons here.

Multilevel data are typically quite constrained at the group level. In the case of TSCS data, this constraint is often binding at either the unit or time levels, especially in cases where there is an extensive hierarchy. Raudenbush and Bryk (2002, 14) write "in the case of hierarchical models, the number of higher-level units ... is usually key in determining whether these large-sample properties (of ML) will apply ... Bayesian methods provide a sensible alternative approach in these cases. SEs will tend to be more realistic than under ML." Further, Bayesian estimation has been shown to have better model fit as data become increasingly hierarchical (Browne and Draper 2006). This is relevant for more complicated TSCS setups, where subunits (such as congressional districts or individuals) are nested in larger units (like states or countries), and both levels have observations across time.

Most MLMs are estimated with either ML or restricted maximum likelihood (REML). In general, the use of ML or REML causes few problems, but as some analysts have noted, ML and REML can suffer from a serious deficiency when used with multilevel data. Both rely on point estimates of the elements in the variance-covariance matrix for inferences. This should come as no surprise since all ML estimators use point estimates of the variance-covariance matrix for confidence intervals and hypothesis tests. For MLMs, however, using such point estimates can be problematic. Gelman and Hill (2006) note that for more complicated MLMs, there may not be enough information in the data to precisely estimate the elements of the variance-covariance matrix.

Then there is the additional factor of unbalanced data, where the number of observations varies per group. We simulate using balanced data in this paper, but TSCS data are often unbalanced. Raudenbush and Bryk (2002) demonstrate that the sampling distribution

[7] See Gelman et al. (2003) for introduction to this methodology, and Jackman (2000, 2004) for a political science–focused introduction.

for the elements of the variance-covariance matrix is skewed to an unknown degree when the data are unbalanced. With balanced data, the sampling distributions are exactly *t*-distributed, but for unbalanced data the sampling distribution is unknown. As Raudenbush and Bryk (2002, 281) note, "... in the unbalanced case the SE estimates for the FE are too small, and hypothesis tests ... will be too liberal."

Both Gelman and Hill (2006) and Raudenbush and Bryk (2002) contend that a Bayesian estimator that averages over the model uncertainty does not suffer from this malady. Because the Bayesian estimator averages over the uncertainty for all the parameters in the model, the researcher's inferences are no longer conditional on the specific point estimates of the variance-covariance matrix. Therefore, inferences are based on the posterior distribution given only the data (Raudenbush and Bryk 2002).

Bayesian models allow the flexibility for researchers to incorporate priors of varying informativeness into their models. Where a body of data and received wisdom exist—as it frequently does in, say, comparative research—it may be appropriate to incorporate different types of informative priors.[8] A major advantage of adopting informative priors is to make explicit the assumptions and existing knowledge used in a given model, information that is too often hidden or implicit in classical approaches.[9]

In our simulations, we use noninformative priors and hyperpriors, as appropriate. This is because we wish to see how the BML estimator performs with a wide variety of sample data under conditions of ignorance about the true values we are trying to recover. The capacity for a researcher to decide to use priors—or not—shows that the Bayesian approach generalizes other inference techniques.

One significant downside of the Bayesian multilevel approach is that—at least in the current implementation in *WinBugs*—it can be computationally intensive, relative to ML algorithms like *lmer()*. This is an important concern because scholars typically test many types of model, and an increase in the cost of testing multiple models could lead to less models being fit. However, with the continual improvements in software and computing power and the declining cost of hardware, this is becoming less of a worry.

2 The Model

We offer multilevel modeling as a solution to issues that commonly occur in TSCS data. Although others have applied MLMs to data sets that are time-series cross-sectional in structure (Western 1998; Shor et al. 2003), we simulate the data to judge how a MLM performs compared to other models. In equation (5), we write a very general MLM that could be used to estimate a TSCS data structure:

$$y_{jt} = \alpha_j + \gamma_t + \beta x_{jt} + \varepsilon_{jt}, \tag{5}$$

where y_{jt} is the outcome variable subscripted for units (j) and time (t). The terms α_j and γ_t are varying-intercept parameters for units and time, respectively.[10] βx_{jt} is a coefficient and

[8] Probably because Bayesian inference is so new in political science, we only have a handful of instances of the use of informative priors, though some do exist (Western and Jackman 1994; Jackman 2004; Gill and Walker 2005). All these papers explore the posterior consequences of different decisions about priors.

[9] For example, MLE assumes complete prior ignorance, an assumption nowhere stated explicitly (Jackman 2004). Other decisions that are in effect made on the basis of prior information include transformations, coding decisions, and unreported models (Western 1998).

[10] Usefully, Bayesian methods allow us to estimate varying intercepts for all the units in the sample. By assigning a common distribution to the set of indicators, Bayesians provide an identifying assumption that makes it unnecessary to drop what are often arbitrary base categories. In the classical framework, multiple intercepts (often implemented with dummy variables) are interpreted as offsets from the dropped-group intercept. A more straightforward method is to estimate the coefficients for all the indicators.

a single predictor that varies across time and space. Finally, ε_{jt} is the level one error. The inclusion of γ_t in the model is one component that is unique to the MLM for TSCS. In the multilevel modeling context, modeling of γ_t takes into account possible contemporaneous correlation that affects all cross-sectional units in a given time period. In more standard models for TSCS (OLS with PCSEs, for example), contemporaneous correlation is viewed as a nuisance property of the error term. As with more standard approaches to TSCS data, $\varepsilon_{jt} \sim N(0, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is an $NT \times NT$ covariance matrix of the errors, with typical elements $(\varepsilon_{it}, \varepsilon_{js})$. To take account of the variability of the errors across units (panel heterogeneity), the elements of $\mathbf{\Omega}$ are estimated for each time period.

The varying intercepts for time and units are allowed as both a systematic and a random component, such that means and variances for each are estimated, allowing for the partial pooling described previously. The model for these intercepts are shown in equations (6) and (7) below:

$$\alpha_j = \alpha_0 + \eta_j, \tag{6}$$

$$\text{where } \eta_j \sim N(0, \sigma_\eta^2).$$

$$\gamma_t = \gamma_0 + \nu_t, \tag{7}$$

$$\text{where } \nu_t \sim N(0, \sigma_\nu^2).$$

The reader should note that unlike ε_{jt}, we assume that the random effects for these varying intercepts are independent across time and space. Moreover, unlike many MLMs, we do not treat the parameter β as random. One could easily relax this assumption and estimate a random coefficient model. Such a model is beyond the confines of this project, but see Beck and Katz (forthcoming) for an example. A substitution of equations (6) and (7) into equation (5) demonstrates the incorporation of errors at different levels into one equation. This is what makes the model multilevel. The conditional estimation used to generate the unit- and time-level error variances, as well as other parameters, is accomplished via Bayesian estimation.

3 Nuisances in TSCS Data

TSCS data typically exhibit nonspherical errors that make OLS regressions problematic. These stem from (a) serial correlation, (b) heteroscedasticity, and (c) contemporaneous correlation. First, because the units are measured over time, the errors can be serially correlated. Second, the assumption of homoscedastic (equal) error variances is often violated. The source of this is usually unit heterogeneity. Third, errors can violate the assumption of cross-sectional independence (Stimson 1985; Beck and Katz 1995).

We have chosen not to generate data with serial correlation. This was done because the OLS/PCSE and FE estimators are not in themselves a correction for serial correlation. They necessitate an additional specification like a lagged dependent variable, a Cochrane-Orcutt/Prais-Winsten transformation, or a linear time trend. Like these methods, multilevel modeling does not by itself correct for serial correlation and the same corrections must be applied. We review heteroscedasticity and contemporaneous correlation in more depth below.

3.1 *Heteroscedasticity*

Unit heterogeneity leads to violations of the homoscedastic error assumptions. In TSCS data, we should expect quite a bit of heterogeneity across units and, perhaps, across time.

Formally, unit heteroscedasticity occurs when the error variance for some units are unequal to the error variances for other units.

It is well known that heteroscedasticity in such data can severely damage uncertainty estimates. Whereas efficiency is therefore an issue, consistency is not. We can accept OLS point estimates, but we must adjust estimates of SEs. Beck-Katz PCSEs are one such method of doing so. Using simulated data, we compare various models to see how each responds to varying levels of heteroscedasticity.

Our MLM approach addresses the issue of unequal variances in two ways. First, we employ varying intercepts for both N and T to account for unit heterogeneity. Second, we explicitly allow the error variance for the group effects to be estimated separately for each group. This helps to account for the fact that some units (and perhaps some time periods) will have higher error variances than others. The flexibility of specifying these group-level variances is characteristic of MLMs (Gelman and Hill 2006).

Heteroscedasticity is treated quite differently in the Bayesian context relative to ML. Since our estimate of uncertainty is derived from the distribution of the posterior, we only have to worry about appropriately modeling the process to measure that distribution accurately. In an ML-based estimator, we would reweight the SEs based on group size, but within a Bayesian context the inferences on each parameter fully takes into account the uncertainty of every other parameter. So long as the heterogeneity is included in the model, one usually ignores the idea of heteroscedasticity.

3.2 *Contemporaneous Correlation*

Contemporaneous correlation refers to the situation when model errors are correlated across space. For example, different units within a single time period would be correlated. This could happen if the units experience a common shock in that period.

When contemporaneous correlation is left alone, it violates a key normality assumption. As Beck (2001) notes, traditional approaches to solving spatial correlation can include fixes like spatial lags. However, PCSEs can adequately clean up this form of correlation as well, as long as the spatial correlation does not function with a lag across time. In the multilevel context, we can deal with contemporaneous correlation by directly modeling varying time intercepts. In doing so, we take into account "time shocks" that operate simultaneously on all cross-sectional units.

4 Monte Carlo Analysis

4.1 *The Data-Generating Process*

For the simulation study, we generated multiple data sets according to the following DGP:[11]

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}, \tag{8}$$

y_{it} is the outcome variable subscripted for units (i) and time (t). The unit effect α_i is drawn from a uniform distribution with a scaling parameter that allows us to vary the size of the unit effect. The covariance matrix of ε_{it}, $\mathbf{\Omega}$, can be manipulated in a number of ways,

[11]The DGP was not created with a multilevel structure in mind. Rather, it was written for an earlier paper in the literature (Kristensen and Wawro 2003) in *Gauss*, then translated into *R* for our purposes.

which we outline in the following section. Here, we do not allow the unit effects to be correlated with x_{it} in our experiments. We set β to 10 so that the estimated coefficient is roughly twice its SE.

4.2 *Estimators*

We use Monte Carlo simulations to compare the performance of least-squares–based estimators for TSCS data with a BML. We do this because most political scientists use some form of least squares with a post-estimation fix for the SEs.

We perform Monte Carlo trials for various levels of N, T, heteroscedasticity, contemporaneous correlation, and unit-effect scalings.[12] We designed the experiments so that the error structure of equation (8) was both heteroskedastic and contemporaneously correlated. For each value of t, we drew each N vector x_{it} from a zero-mean normal distribution. We simulated heteroskedasticity by setting the variance of the first $N/2$ units to 1, while we manipulated the variances of the second set of units. Next, we constructed the covariance matrix so that all pairs of units share a common correlation that we can manipulate in the analysis. The variances and covariances of the errors were proportional to the corresponding variances and covariances of x_{it}. This design allows for both panel heteroskedasticity and contemporaneous correlation serious enough to cause incorrect estimates of the SEs, if not taken into account.

The standard Monte Carlo procedure is to perform some large number of trials—typically 1000 or more. However, performing Monte Carlo simulations with Bayesian models is extremely time consuming as the same experiment may take anywhere from 100 to 300 *times* as long as classical estimators on the same computer.[13] For the results reported here, only 250 Monte Carlo trials were performed. We find that larger number of simulations do not affect our results.

4.2.1 Classical Estimators

We ran the classical OLS and FE estimators.[14] We pair the former with regular OLS SEs and (Beck and Katz 1995) PCSEs.[15] The latter are paired with regular FE SEs and FE robust SEs (Arellano 1987).

4.2.2 Bayesian Multilevel Model

We ran the BML with two chains iterating 2100 times. We discarded half the iterations as a burn-in and thinned the chains by three to save memory. This generated posterior distributions of length 700. Both chains converged for all estimated parameters, according to the Gelman-Rubin $\hat{R}$ diagnostic (Gelman et al. 2003).

[12] Unit-effect scalings, abbreviated as A.Sc in the Results, allow us to simulate the extremes of no and massive unit effects, as well as intermediate levels.

[13] One reason for this is that the Bayesian model here was estimated with *WinBugs* using a Markov Chain Monte Carlo algorithm. So our results are essentially 250 Monte Carlo trials, each consisting of 2100 Monte Carlo iterations!

[14] The FE allow for varying intercepts for units.

[15] Perhaps because of the wide popularity of the PCSE approach, many scholars erroneously use OLS with PCSEs for TSCS data without accounting for potential unit effects. Poor performance by OLS with PCSEs in these scenarios is therefore the fault of researchers, not the estimator itself. We seek to show the consequences of such a common misspecification below.

5 Results

To benchmark the various estimators, we report several test statistics: bias, root mean–squared error (RMSE), relative efficiency, and optimism. We also display histograms of the distributions of the various β's and SEs estimated by the various models.

5.1 *Bias*

First, we calculate the bias of the estimators calculated as shown in equation (9) for the simulated β's. The closer this value is to 100, the less biased, on average, is the estimator.

$$\text{Bias} = 100 \cdot (\overline{\beta}/\beta_{\text{true}}). \tag{9}$$

None of the estimators is expected to be biased, and Monte Carlo results (not shown) confirm this result. Differences between them are found in the SEs, not in the coefficients.

5.2 *RMSE and Efficiency*

One way to capture the bias and the efficiency of the estimators simultaneously is to examine the distribution of the β's around the true β; this is easily done with the RMSE. Equation (10) shows that estimators are penalized for being biased, as well as having a larger spread around the true β.

$$\text{RMSE} = \sqrt{\frac{\sum_{l=1}^{n\text{sims}} (\beta^{(l)} - \beta_{\text{true}})^2}{n}}. \tag{10}$$

The relative efficiency of the BML estimator over that of the OLS and FE estimators can be derived by calculating ratios of RMSE scores, shown in equation (11). This indicates the degree to which the BML estimator's RMSE score is smaller than that of another estimator.

$$\text{Relative Efficiency} = 100 \cdot \frac{\text{RMSE}_{\text{OLS|FE}}}{\text{RMSE}_{\text{BML}}}. \tag{11}$$

Figures 1 and 2 graphically show the RMSE and relative efficiency results found in Table 1 for data with varying amounts of *N*, *T*, contemporaneous correlation, heteroscedasticity, and the scale of the unit effects. These results show that the BML estimator is more efficient than classical estimators for the DGPs we have run in this paper.

This efficiency advantage is also clearly seen in Fig. 3 that shows the histograms of the 250 β's estimated by each model for condition 2 (see row 2 of any table). The estimated BML β's cluster far more tightly (95% interval is 9.6–10.4) around the true value of 10 than either OLS or FE.

Fig. 4 shows another view of this; the reported BML SEs are quite a bit smaller than they are for the other classical SEs. The BML SEs cluster around 0.16–0.25, whereas PCSEs cluster near 0.34–0.81, and FE robust SEs cluster around 0.16–0.43.

5.3 *Optimism*

In the previous section, we have seen that the reported BML SEs are smaller than those of the other classical SEs. But are those smaller errors justified? To characterize the uncertainty estimates returned by the estimators for the simulated betas, we calibrate them

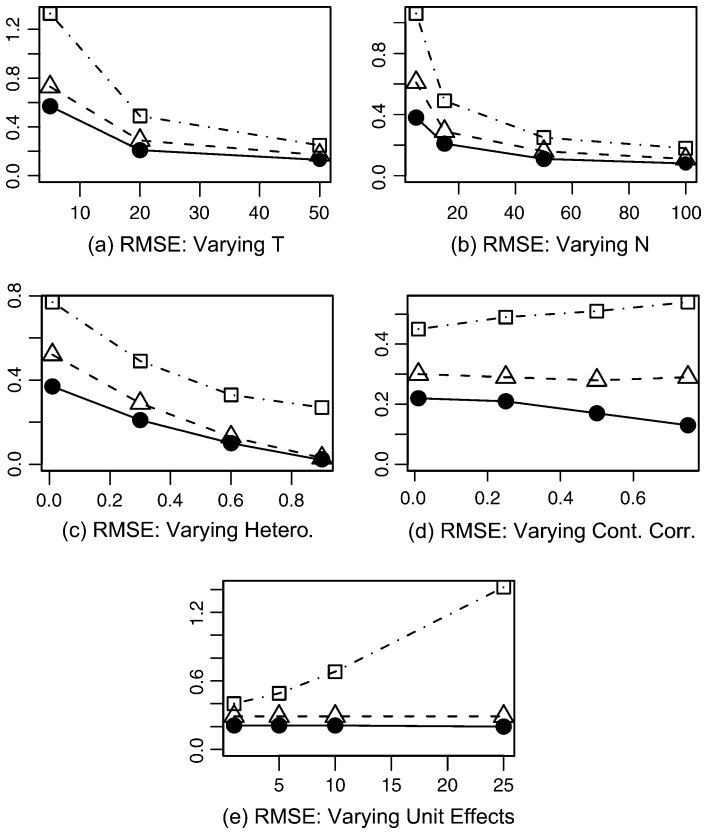

Fig. 1 Calculated RMSE statistics for OLS (dash-dot), FE (dash), and BML (solid) estimators, under varying levels of *T*, *N*, heteroscedasticity, contemporaneous correlation, and the scaling of unit effects.

by the spread of the estimated β's. This statistic is called "overconfidence" by Beck and Katz (1995) and "optimism" by Kristensen and Wawro (2003). Values over 100 indicate that the true sampling variability of an estimator is larger than the reported estimate of variability, indicating that SEs are "too confident." Conversely, optimism values under 100 indicate that the SEs are "underconfident" (too modest). Equation (12) shows the calculation for optimism.

$$\text{Optimism} = 100 \cdot \frac{\sqrt{\sum_{l=1}^{n\text{sims}} (\beta^{(l)} - \overline{\beta})^2}}{\sqrt{\sum_{l=1}^{n\text{sims}} (\text{SE}(\beta^{(l)}))^2}}. \tag{12}$$

Table 2 and Fig. 5 show the results of equation (12) applied to various conditions simulated by the DGP. It can be seen that the reported uncertainty estimates for the MLM are more characteristic of the true variance of the estimator than the even more uncertain estimates of PCSEs and robust SEs, not to mention the uncorrected SEs.

On the basis of these MC results, OLS SEs and FE SEs can be said to be improperly overmodest. PCSEs are slightly too modest, as well, but to a far smaller degree. In other words, the distribution of the estimated β's is tighter than we would expect on the basis of

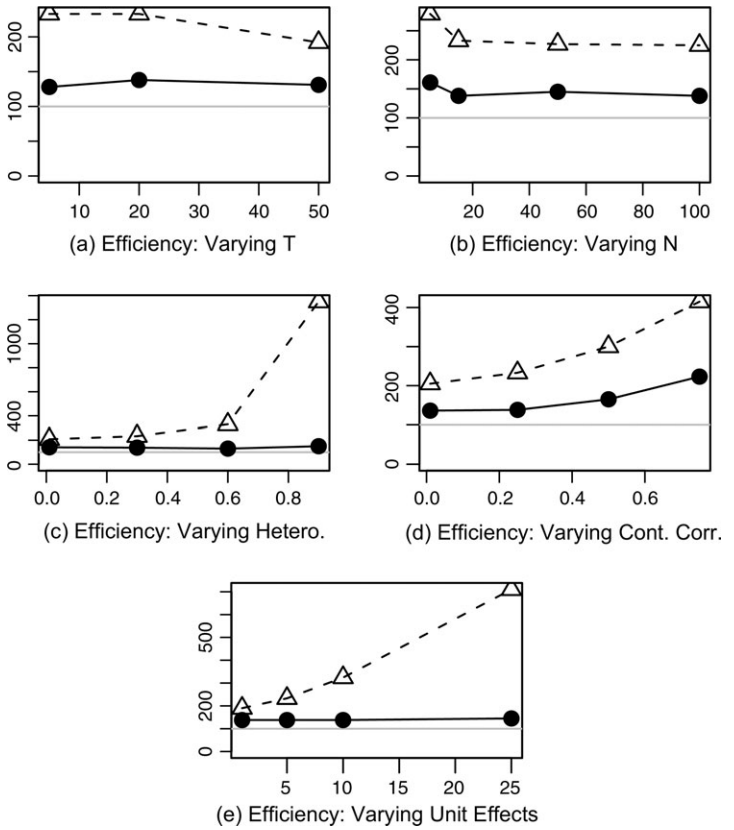

Fig. 2 Calculated relative efficiency advantage for the BML over OLS (dash) and FE (solid), under varying levels of *T*, *N*, heteroscedasticity, contemporaneous correlation, and the scaling of unit effects.

reported SEs. The consequence of this is that social scientists might be led into making a type II error: accepting a null that should be rejected. This inferential error is perhaps less serious than making a type I error, but is nevertheless to be avoided. This finding is especially true when the nuisances of TSCS data are very powerful.

6 The MLM Generalized

We have considered MLMs with varying intercepts. However, we can easily extend the scope of these models to include varying slopes as well. This would allow us to estimate different causal effects of a substantive predictor of interest for different units and/or time periods. This is often called, perhaps imprecisely, a "random coefficient model."[16]

We have also not considered the prospect of group-level predictors, that is, covariates at an aggregate level of analysis. Using FE under the classical paradigm may tell us that cross-sectional units differ, but not why. The multilevel approach allows us to easily include group-level predictors to explain variation in intercepts and/or slopes across units or time. If the individual level of analysis is the state-year, then we can have state-level predictors that are constant over time (perhaps state area), or year-level predictors that

[16]Our varying-intercept model—as are all Bayesian models—is a random coefficient model because all estimated parameters come from random posterior distributions.

Table 1 RMSE/relative efficiency results

	N	*T*	*Cor*[a]	*Het*[b]	*A.Sc*[c]	*OLS RMSE*	*FE RMSE*	*BML RMSE*	*OLS/BML Eff*	*FE/BML Eff*
1	15	5	0.25	0.3	5	1.33	0.73	0.57	233	128
2	15	20	0.25	0.3	5	0.49	0.29	0.21	233	138
3	15	50	0.25	0.3	5	0.25	0.17	0.13	192	131
4	5	20	0.25	0.3	5	1.06	0.61	0.38	279	161
5	50	20	0.25	0.3	5	0.25	0.16	0.11	227	145
6	500	20	0.25	0.3	5	0.18	0.11	0.08	225	138
7	15	20	0.25	0.0	5	0.77	0.52	0.37	208	141
8	15	20	0.25	0.6	5	0.33	0.13	0.10	330	130
9	15	20	0.25	0.9	5	0.27	0.03	0.02	1350	150
10	15	20	0.01	0.3	5	0.45	0.30	0.22	205	136
11	15	20	0.50	0.3	5	0.51	0.28	0.17	300	165
12	15	20	0.75	0.3	5	0.54	0.29	0.13	415	223
13	15	20	0.25	0.3	1	0.40	0.29	0.21	190	138
14	15	20	0.25	0.3	10	0.68	0.29	0.21	324	138
15	15	20	0.25	0.3	25	1.42	0.29	0.20	710	145

[a]Contemporaneous correlation.
[b]SD of $1/\sigma$ normalized.
[c]Scale parameter of unit effects.

are constant over space (perhaps national gross domestic product). This allows us to better estimate the variance of group-level units from the data. Including these group-level predictors would be trivial in a MLM, and would significantly improve model fit.

Such nested covariates can also aid when individual-level predictors and groups correlate. Such a correlation has often led researchers to avoid models with varying intercepts

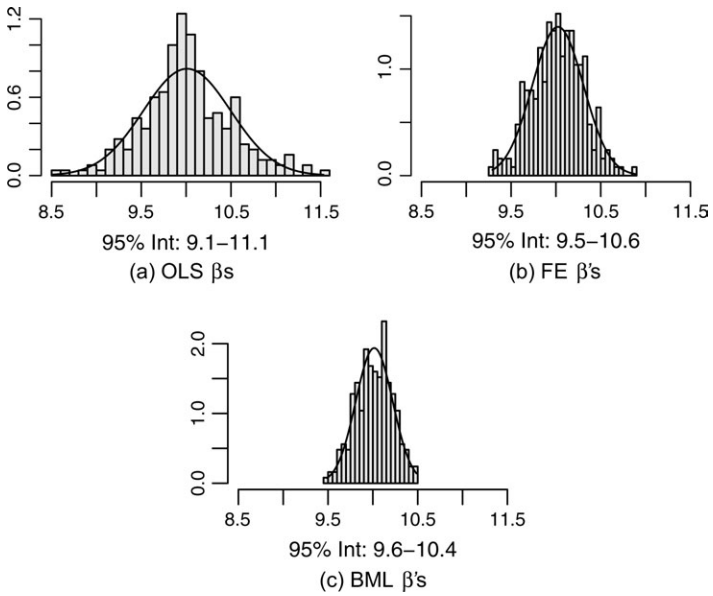

Fig. 3 Histogram of estimated β's for OLS, FE, and BLMs. *N*(sims) = 250, $N = 15$, $T = 20$, Cor = 0.25, Het = 0.3, A.Scale = 5.

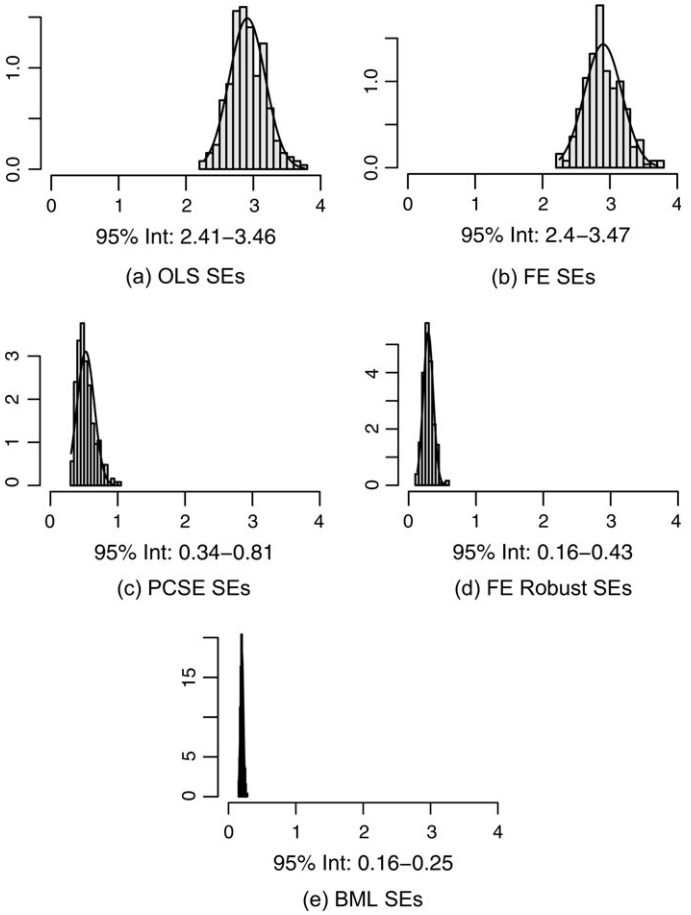

Fig. 4 Histogram of estimated SEs for OLS, FE, and BMLs. N(sims) = 250, N = 15, T = 20, Cor = 0.25, Het = 0.3, A.Scale = 5.

(or slopes) modeled with error because of potential bias or poor uncertainty estimates. Bafumi and Gelman (2006) show that this problem can easily be resolved by calculating the mean per group of the correlating predictor and including it as a covariate at the group level. Classical estimators with unit or time effects would simply drop sluggish or invariant unit-level or time-level predictors, or bias them downwards toward zero.[17]

Finally, MLMs can incorporate additional levels of hierarchy or structure. For example, a TSCS data set of congressional districts might incorporate not only district- and time-invariant predictors but also predictors at the state and national levels (Shor 2004). We would expect these to improve model fit and estimates of uncertainty, as well as be more theoretically justified.

[17]One might drop additional indicators or employ a two-stage approach to estimate slow or nonvarying group-level predictors. As compared to the multilevel modeling described here, however, these approaches are less intuitive and more complicated (for example, they require additional post-estimation fixes such as correcting SEs).

Table 2 Optimism results

	N	*T*	*Cor*[a]	*Het*[b]	*A.Sc*[c]	*OLS Opt*	*FE Opt*	*PCSE Opt*	*FE Robust Opt*	*BML Opt*
1	15	5	0.25	0.3	5	22	12	81	105	108
2	15	20	0.25	0.3	5	17	10	90	95	102
3	15	50	0.25	0.3	5	13	9	85	96	104
4	5	20	0.25	0.3	5	21	12	95	104	92
5	50	20	0.25	0.3	5	16	10	90	101	105
6	500	20	0.25	0.3	5	16	10	95	102	107
7	15	20	0.25	0.0	5	26	18	94	95	102
8	15	20	0.25	0.6	5	11	5	81	95	102
9	15	20	0.25	0.9	5	9	1	73	95	101
10	15	20	0.01	0.3	5	15	10	95	100	102
11	15	20	0.50	0.3	5	18	10	86	94	102
12	15	20	0.75	0.3	5	19	10	82	94	101
13	15	20	0.25	0.3	1	14	10	87	95	102
14	15	20	0.25	0.3	10	23	10	93	95	102
15	15	20	0.25	0.3	25	44	10	95	95	102

[a]Contemporaneous correlation.
[b]SD of $1/\sigma$ normalized.
[c]Scale parameter of unit effects.

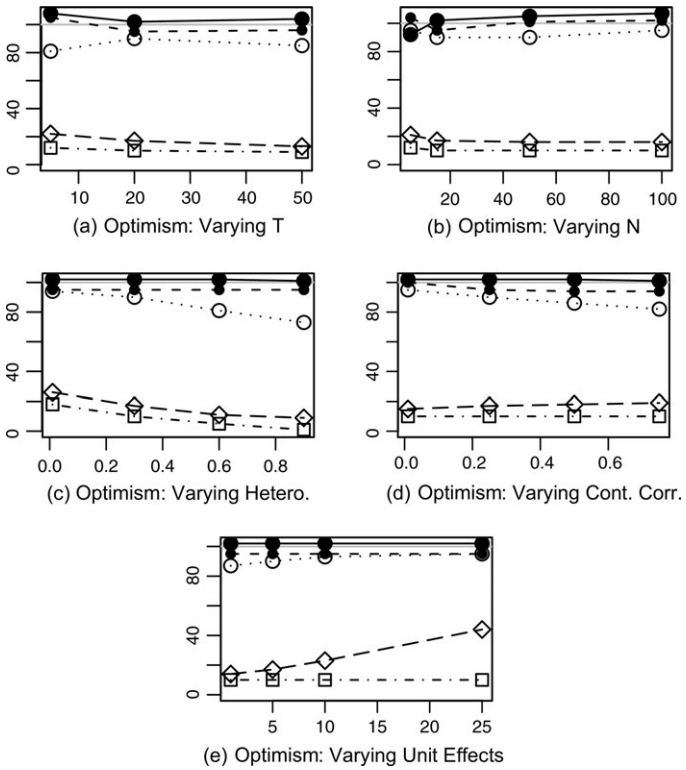

Fig. 5 Calculated optimism statistics for OLS SE (dash and triangle), PCSE (dot), FE SE (dash-dot), FE Robust SE (dark short dash and empty circle), and BML SE (solid) under varying levels of *T*, *N*, heteroscedasticity, contemporaneous correlation, and the scaling of unit effects.

7 Conclusion

We have shown that a MLM can adequately address two characteristic problems of TSCS data: heteroscedasticity and contemporaneous correlation. It does this at least as well, and often quite a bit better, than other common estimators for such data.

All the tested estimators are—on average—unbiased. However, the spread of the estimated β's is wider for OLS and the FE estimator. The multilevel estimator is substantially more efficient than either the OLS and FE estimators for the simulated TSCS data. This efficiency advantage can increase dramatically under aggressively problematic levels of heteroscedasticity and contemporaneous correlation. The quality of the reported SEs is better than classical OLS SEs, FE SEs, PCSEs, and FE robust SEs. They are dramatically better than the first two. Researchers can feel confident about employing this estimator for their TSCS data sets.

The multilevel approach offers some unique advantages over classical estimators for TSCS data. Limited data can be compensated for with partial pooling. Fully or partially time-invariant predictors can be estimated simultaneously with varying group-level intercepts. MLMs are easily extended to include additional levels of analysis. Unlike ML, priors can be incorporated—or not—at the researcher's discretion. In sum, MLMs are both more general and flexible than their classical counterparts.

References

Arellano, Manuel. 1987. Computing robust standard errors for within-group estimators. *Oxford Bulletin of Economics and Statistics* 49:431–4.

Bafumi, Joseph. 2004a. The macro micro link in vote choice models: A Bayesian multilevel approach. Presented at the 2004 annual meeting of the Midwest Political Science Association, Chicago, IL.

———. 2004b. The stubborn American voter. Presented at the 2004 annual meeting of the American Political Science Association, Chicago, IL.

Bafumi, Joseph, and Andrew Gelman. 2006. Fitting multilevel models when predictors and group effects correlate. Paper presented at the annual meeting of the American Political Science Association, Philadelphia, PA.

Bafumi, Joseph, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 13:171–87.

Bartels, Larry M. 1996. Pooling disparate observations. *American Journal of Political Science* 40:905–42.

Beck, Nathaniel. 2001. Time-series-cross-section data: What have we learned in the past few years. *Annual Review of Political Science* 4:271–93.

Beck, Nathaniel, and Jonathan N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89:634–47.

Beck, Nathaniel, and Jonathan N. Katz. 2007. Random coefficient models for time-series-cross-section data: Monte Carlo experiments. *Political Analysis.* 10.1093/pan/mpl001.

Browne, William J., and David Draper. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* 1:473–514.

Gelman, Andrew. 2006. Multilevel (hierarchical) modeling: What it can and can't do. *Technometrics* 48:432–5.

Gelman, Andrew, John S. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian data analysis.* 2nd ed. Boca Raton, FL: Chapman and Hall.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gelman, Andrew, and Gary King. 1993. Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* 23:409–51.

Gelman, Andrew, Boris Shor, Joseph Bafumi, and David Park. Forthcoming. Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science.*

Gill, Jeff, and Lee D. Walker. 2005. Elicited priors for Bayesian model specifications in political science research. *The Journal of Politics* 67:841–72.

Jackman, Simon. 2000. Estimation and inference are missing data problems: Unifying social science statistics via Bayesian simulation. *Political Analysis* 8:307–32.

———. 2004. Bayesian analysis for political research. *Annual Review of Political Science* 7:483–505.

Kristensen, Ida P., and Gregory Wawro. 2003. Lagging the dog? The robustness of panel corrected standard errors in the presence of serial correlation and observation specific effects. Presented at the annual meeting of the Political Methodology Conference, Minneapolis, MN.

Park, David. 2001. Representation in the American states: The 100th Senate and their electorate. Working paper.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* 12:375–85.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods.* 2nd ed. Newbury Park, CA: Sage Publications.

Shor, Boris. 2004. Taking context into account: Testing partisan, institutional, and electoral theories of political influence on defense procurement in congressional districts, 98th–102nd congress. Paper presented at the annual meeting of the American Political Science Association, Chicago, IL.

———. 2006. A Bayesian multilevel model of federal spending, 1983–2001. Working paper.

Shor, Boris, Joseph Bafumi, David K. Park, and Andrew Gelman. 2003. Multilevel modeling of time series data. Paper presented at the annual meeting of the American Political Science Association, Philadelphia, PA.

Steenbergen, Marco R., and Bradford S. Jones. 2002. Modeling multilevel data structures. *American Journal of Political Science* 46:218–37.

Stimson, James A. 1985. Regression models in space and time: A statistical essay. *American Journal of Political Science* 29:914–47.

Western, Bruce. 1998. Causal heterogeneity in comparative research: A Bayesian modelling approach. *American Journal of Political Science* 42:1233–59.

Western, Bruce, and Simon Jackman. 1994. Bayesian inference for comparative research. *American Political Science Review* 88:412–33.