Advance Access publication June 27, 2006 *Political Analysis* (2007) 15:182–195
doi:10.1093/pan/mpl001

Random Coefficient Models for Time-Series–Cross-Section Data: Monte Carlo Experiments

Nathaniel Beck
Department of Politics, New York University, New York, NY 10003
e-mail: nathaniel.beck@nyu.edu (corresponding author)

Jonathan N. Katz
Division of the Humanities and Social Sciences,
California Institute of Technology, Pasadena, CA 91125
e-mail: jkatz@caltech.edu

This article considers random coefficient models (RCMs) for time-series–cross-section data. These models allow for unit to unit variation in the model parameters. The heart of the article compares the finite sample properties of the fully pooled estimator, the unit by unit (unpooled) estimator, and the (maximum likelihood) RCM estimator. The maximum likelihood estimator RCM performs well, even where the data were generated so that the RCM would be problematic. In an appendix, we show that the most common feasible generalized least squares estimator of the RCM models is always inferior to the maximum likelihood estimator, and in smaller samples dramatically so.

1 Random Coefficient Models and Time-Series–Cross-Section Data

The use of time-series–cross-section (TSCS) data has become very common in the study of comparative political economy.[1] TSCS data consist of time-series observations at regular intervals on a fixed set of units. The paradigmatic comparative political economy studies (e.g., Garrett 1998) have about 30 annual observations on about 20 advanced industrial democracies; the dependent variable in these studies is typically a measure of economic policy or outcome, with the independent variables being the economic and political determinants of those outcomes or policies. Whereas both the number of units and the length of the time series vary from study to study, the numbers in the Garrett study

Authors' note: We gratefully acknowledge the financial support of the National Science Foundation. Katz also acknowledges the support of the Center for Advanced Study in the Behavioral Sciences. We are thankful to Jake Bowers, Rob Franzese, Andy Gelman, Sandy Gordon, Bill Greene, and Luke Keele for comments; to Larry Bartels for always reminding us that our judgment may outperform the data; as well as to the anonymous reviewers of Political Analysis.

[1]We use the term comparative political economy instead of comparative politics since there have recently been a series of applications in comparative political behavior that use methods similar to those investigated here. Those applications are, as we shall see, to data sets that do not look like typical TSCS data sets, and so we wish to keep the two types of applications distinct.

are quite typical; TSCS typically have between 10 and 50 observations on between 10 and 50 units.[2]

We distinguish TSCS data from both panel data and other types of multilevel or hierarchical data. The former have many units (typically a thousand or more survey respondents) observed for a very small number of "waves" (as few as 2, almost always fewer than 10). Thus, asymptotic results (in the number of units) give good insight into the properties of estimators for panel data but provide much less guidance for estimation of models using TSCS data. In multilevel studies that are common in education, there are typically few observations on many units, with a common data set being many schools each containing few classrooms. Again, statistical results that depend on the properties of averaging over a large number of units are informative in this context. Alternatively, recent multilevel work in comparative political behavior (Kedar and Shively 2005; Steenbergen and Jones 2002) has a large number of survey respondents observed in a few countries. Here we get good guidance from asymptotic results that depend on a large number of observations per unit but do not depend on their being a large number of those units. In TSCS data, we have neither a large number of observations per unit nor a large number of units, and so asymptotic results in either the number of observations per unit or the number of units may not be helpful. It is for this reason that we turn to Monte Carlo simulations to analyze the finite sample properties of one type of TSCS estimator using simulated data that mimic the types of TSCS data seen in the study of comparative political economy. In particular, we consider estimators designed to deal with issues related to parameter variation across, but not within, the units.

In general, TSCS analysts seem willing to assume that all countries (units) are completely homogeneous, such that a fully pooled model is appropriate. This assumption is often made without considering the alternatives. There are, of course, such alternatives. Many analysts allow for unit-specific intercepts, that is, fixed effects. But there are relatively few attempts to go beyond this limited heterogeneity. Obviously, we must assume enough homogeneity to allow for estimation; if every observation is unique, we can do no science. But it is not necessary to assume that the only alternative to complete uniqueness is complete pooling (or its close cousin, pooling other than for the unit-specific intercepts). At first glance, the commitment to homogeneity is a bit odd since a model that allows for heterogeneity, the random coefficient model (RCM), has been known under various names (hierarchical, mixed, multilevel, random coefficient, and varying parameter models, at least) for over half a century. Such models were considered in the light of comparative political economy by Western (1998). Should TSCS analysts routinely entertain the RCM? Does it work well for the kinds of data and the kinds of questions typically seen in comparative political economy? In this article, we continue the investigation of this question.

We begin by noting that the issue of whether to pool or not or, more accurately, how much to pool confronts every researcher. In an important, if insufficiently utilized piece, Bartels (1996) argues that we are always in the position of deciding how much we should pool some observations with others and we always have a choice ranging from complete pooling to assuming that the data have nothing to do with each other. He notes that, in general, political scientists seem to assume that either data completely pool or that some data are completely irrelevant, ignoring the in between position. The solution that Bartels proposes is that one should estimate a model allowing for varying degrees of pooling and then make a scientific decision after examining the locus of all such estimates. The

[2]Adolph, Butler, and Wilson (2005, 4) have a graph of N and T for over 100 TSCS studies, with the large bulk of the studies falling within the range given here.

procedure involves much judgment since Bartels works in a purely cross-sectional context; in that context, the data alone can never determine the appropriate degree of pooling.

The situation is much simpler when data are grouped by unit and where we can assume that parameters vary between but not within those units. This leads to the well-known hierarchical (or multilevel or mixed) model; these are the models we examine here. For TSCS data, the most common nomenclature for such models is the RCM, and we use this nomenclature here, noting that regardless of nomenclature the various models are formally identical. But since the various models are typically used in very different research situations (with very different numbers of units and observations per unit), these identical methods may perform differently in those different research situations.

Western has described the RCM in a Bayesian context. Although his work is reasonably well cited, we have found precious few (if any) applications of Western's method to substantive issues in comparative politics.[3] In this article we show that the RCM, estimated via classical maximum likelihood, performs very well and should be more utilized by students of comparative political economy. Since the particular implementation we examine, that of Pinheiro and Bates (2000), is easily accessible in the popular software package R, there is no practical impediment to use of RCMs. We conjecture that at least some of the lack of interest in RCMs was due to the view that they were simply not worth the extra effort to estimate, a view that is clearly incorrect today.

In this article, our interest is in examining the performance of the RCM for what we consider to be typical TSCS data. Since the statistical derivations are well known and described elsewhere (with the most relevant being Western [1998], Pinheiro and Bates [2000], and Hsiao [2003]), we omit them here. The heart of the article is a series of Monte Carlo simulations of the finite sample properties of various TSCS estimators. Sections 2 and 3 of the article lay out the notation of the RCM and the various estimators and discuss the Monte Carlo setup. Section 4 provides Monte Carlo evidence on specific estimators and Section 5 concludes. The appendix consider the properties of one the most common estimators of RCMs, the Swamy-Hsiao feasible generalized least squares (FGLS) estimator.

2 Estimating RCMs

We assume standard TSCS data with a continuous dependent variable. The fully pooled model is

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\beta} + \varepsilon_{i,t}, \quad i = 1, \ldots, N, \;\; t = 1, \ldots, T, \tag{1}$$

where $\mathbf{x}_{i,t}$ is a (row) vector of K independent variables for unit i at period t and $\boldsymbol{\beta}$ is a K-vector of parameters. We assume that the errors are serially independent (perhaps after including a lagged dependent variable in the specification[4]). We also assume that $\varepsilon_{i,t} \overset{\text{ind}}{\sim} N(0, \sigma^2)$, so the error process is (for TSCS data) unusually simple. All these simplifying assumptions are made so that we can focus on the issue of parameter heterogeneity. In the real world of research, issues will obviously be more complicated. The simple pooled model can be estimated by ordinary least squares (OLS). Throughout we assume that neither N nor T are large, but neither are they tiny (so asymptotic theory may not well describe the behavior of averages but nor is averaging hopeless).

[3]The only applications we know of are by Bruce Western himself (Western [1997] and Wallerstein and Western [2000]). It is possible that current interest in Bayesian methods will lead to a new interest in RCMs estimated via Bayesian methods.

[4]Our interest in this article is not in dynamics, so the important issue of modeling dynamics need not detain us here.

The less pooled model has

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\beta}_i + \varepsilon_{i,t}. \tag{2}$$

We use the term "less pooled" because the notation allows for the $\boldsymbol{\beta}_i$ to follow a variety of patterns. We could, for instance, allow each unit's $\boldsymbol{\beta}_i$ to be completely unrelated, the fully unpooled model, which would be estimated by unit by unit OLS.

The RCM adjoins to equation (2) the assumption that the $\boldsymbol{\beta}_i$ are all related, in particular they are draws from a multivariate normal distribution. Thus, the RCM is equation (2) with the additional assumption

$$\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Gamma}), \tag{3}$$

where $\boldsymbol{\Gamma}$ is a matrix of variance and covariance terms to be estimated. $\boldsymbol{\Gamma}$ indicates the degree of the heterogeneity of the unit parameters ($\boldsymbol{\Gamma} = \mathbf{0}$ indicates perfect homogeneity). An important (and restrictive) assumption is that the stochastic process which generates the $\boldsymbol{\beta}_i$ is independent of the error process and is also uncorrelated with the vector of independent variables.

For some purposes it is simpler to write

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\nu}_i, \tag{4}$$

$$\boldsymbol{\nu} \sim N(\mathbf{0}, \boldsymbol{\Gamma}) \tag{5}$$

so

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\beta} + \{\mathbf{x}_{i,t}\boldsymbol{\nu}_i + \varepsilon_{i,t}\}. \tag{6}$$

Thus, the RCM can be seen as a linear model with a complicated error term (in braces).

The independence assumptions we make ensure that the $\mathbf{x}_{i,t}$ are uncorrelated with this complicated error term; the complications of the error term imply that the conditions of the Gauss-Markov theorem do not hold. As a consequence, OLS estimation of equation (6) will yield consistent, but inefficient, estimates of $\boldsymbol{\beta}$ with possibly incorrect standard errors.

Although the assumption of parameters being drawn from a distribution is not a natural assumption for a classicist, the RCM can be estimated by classical (maximum likelihood) methods as well as by Bayesian-inspired Markov chain Monte Carlo methods as in Western. Working with a classical interpretation, we use maximum likelihood methods, in particular, those of Pinheiro and Bates (2000), which are implemented in their "nonlinear mixed estimation" (nlme) R package (we use the lme function). Although the maximization is quite complicated, this is "just" maximum likelihood, so we refer readers to the Pinheiro and Bates book for details.[5]

If the data are generated by the RCM process, the RCM estimates of $\boldsymbol{\beta}$ will be more efficient than the corresponding OLS estimates, and the RCM standard errors will be correct. One set of experimental results will be used to assess the relative efficiency of the RCM estimate of $\boldsymbol{\beta}$ as compared to the corresponding OLS estimate for typical TSCS

[5]For those who prefer Stata to R, Stata also implements a maximum likelihood routine that, though less flexible, yields estimates very similar to those from the R package. There are many computational methods for actually maximizing the likelihood. Greene (2003, 512–7), for example, recommends simulated maximum likelihood; Western used Markov chain Monte Carlo. Choosing between the methods of doing maximum likelihood estimation is not our interest here.

data. Efficiency is assessed by examining the root mean squared error (RMSE) of the various estimators, where the errors here measure the deviation of the estimated parameters from the known truth, and the averaging is over the large number of iterations of the simulations.

So far we have discussed estimating the fundamental parameters of the RCM, the overall $\boldsymbol{\beta}$, and the variance of the distribution of the $\boldsymbol{\beta}_i$'s, $\boldsymbol{\Gamma}$. One can clearly estimate the $\boldsymbol{\beta}_i$ using unit by unit OLS. Such estimates will have usual optimal asymptotic properties, and if T were large these would be of interest. But, for typical TSCS T's we might be able to improve on unit by unit OLS. Since with smaller T's the unit by unit OLS estimates may show very high variability, we might choose to use the overall pooled estimate of $\boldsymbol{\beta}$ as the estimate for each of the $\boldsymbol{\beta}_i$; this would trade increased bias for decreased variance. We refer to this as the pooled estimate of $\boldsymbol{\beta}_i$.

The RCM provides an alternative way to estimate the $\boldsymbol{\beta}_i$. This consists of finding the "best linear unbiased predictor" (BLUP), which has $E(\boldsymbol{\beta}_i) = \boldsymbol{\beta}$ and lowest error loss in the class of such linear unbiased estimators (or predictors). The various texts we refer to provide the formulae, but the BLUP can be seen as shrinking back the unit by unit OLS estimates to the overall pooled estimate, with the degree of shrinkage a function of the uncertainty of the estimates.[6] Hence we use the same experiments to compare RCM estimates of the $\boldsymbol{\beta}_i$ to the unit by unit and pooled OLS estimates. As for $\boldsymbol{\beta}$, relative efficiency is assessed via the RMSE of the various estimators over the runs of the simulations.

Finally, we must ask how the RCM performs if the $\boldsymbol{\beta}_i$ vary but do not look like draws from a normal distribution. We thus repeat our experiments but draw the $\boldsymbol{\beta}_i$ from a gamma distribution; this distribution is both asymmetric and has a longer right tail than does the normal. Obviously we could improve on the efficiency of the RCM if we knew that the $\boldsymbol{\beta}_i$ were drawn from a gamma distribution; here the question we ask is how well does the standard (normal) RCM do as compared to OLS when the normality assumption is violated. We also conducted another set of experiments where two of the units were extreme outliers; does the RCM enable us to precisely enough estimate the $\boldsymbol{\beta}_i$ so we can pick up those outliers?

Before turning to the simulations, we note that readers may be puzzled by two apparent omissions. First, we note that comparativists are less interested in the $\boldsymbol{\beta}_i$ *per se* than in explaining variation in the $\boldsymbol{\beta}_i$. Thus, they adjoin to the RCM some unit level variables which help explain the unit to unit variation in the $\boldsymbol{\beta}_i$, yielding

$$\boldsymbol{\beta}_i = \alpha + \mathbf{z}_i\boldsymbol{\kappa} + \mu_i, \tag{7}$$

where the $\mathbf{z}_i$ are covariates that pertain to a unit but do not vary over time (structural characteristics of the country). But substituting equation (7) into equation (2) shows that this is still an RCM (equivalent to equation 6) but one that contains *nonrandom* multiplicative terms interacting the $\mathbf{x}$ and $\mathbf{z}$ covariates. Thus, although modeling the random coefficient as a function of structural covariates should be of great interest to students of comparative politics, this adds nothing new statistically to the RCM and hence we do not consider it further in this article. We stress, however, that the reasoning behind equation (7) is one of the fundamental reasons why the RCM is so valuable for the study of comparative

[6]That the BLUP is superior to the unbiased unit by unit OLS estimates is a remarkable result. As Efron and Morris have shown in a number of papers (e.g., Efron and Morris 1975), the result follows from fundamental work of Charles Stein. A very simple explanation is that the BLUP is a weighted average of a very stable but biased estimator (the estimate of the overall $\boldsymbol{\beta}$) and an unbiased but highly unstable estimator (the OLS estimate of the unit $\boldsymbol{\beta}_i$), with the weighting being determined by the variability of each estimate. The resulting average, although still biased, has lower squared error loss, with the weighting procedure guarding against both introducing too much bias or putting too much weight on what might be a highly variable estimate.

politics. Our simulations can be seen as assessing the gains from including μ_i in equation (7) since without the μ we would just use OLS with interaction terms.

Readers familiar with RCMs will also note that we have omitted a standard RCM estimator that uses FGLS. The FGLS estimator is due to Swamy (1971) though many know this work through Hsiao (2003). This estimator was derived decades ago, long before maximum likelihood estimation of the RCM was computationally practical. It is still implemented in standard statistical packages and is still discussed in current work as an alternative to full maximum likelihood estimation. Thus, for example, Hsiao (2003, 146–7) first presents the FGLS estimator and then states "[a]lternatively, we can use the Bayes model estimator suggested by [Smith]" providing no guidance as to which should be preferred. In the one example he provides, only the FGLS estimator was used.

In Beck and Katz (2004), we showed that the Swamy-Hsiao FGLS estimator can have quite poor finite sample properties. This is because making the estimated variance-covariance matrix positive definite required assuming that the sampling variance of the individual unit estimators was zero, an assumption that could be very bad in practice. Although here we work only with the maximum likelihood estimator, and so need not worry about the FGLS estimator, the latter is still commonly used (as can be inferred from the citation in the previous paragraph). Since the properties of the FGLS estimator are orthogonal to our interests here, we present the major findings from our previous work in the appendix to this article, with full details available in our previous paper. The appendix shows that the maximum likelihood estimator is better, and for smaller T substantially better, than the Swamy-Hsiao FGLS estimator. Given that it is now practical for researchers to do full maximum likelihood estimation of the RCM, there is no reason to continue to consider the FGLS estimator.

3 Design of the Simulations

We designed the simulations to mimic the simplest form of TSCS data so we could compare the RCM and OLS estimators. Thus, the only complications of the data generation process we consider are that model parameters vary from unit to unit. We also choose our sample sizes to mimic those typically found in TSCS data. Since we have found that results do not vary much with N (at least for plausible TSCS values of N), we fixed N at 20. We then varied T from 5 to 50 in increments of five.

We work with the simplest of models (returning in the conclusion to speculate on what might happen in more complex situations). Thus, our model has only single covariate, $x_{i,t}$. We always draw from zero mean distributions but estimate a single constant term (that is, in truth there are neither unit effects nor a constant term, but we estimate the constant term though no effects).

The simulations consist of 1000 replications. Before the start of the replications, we drew the $N \times T$ regressors from $x_{i,t} \overset{\text{ind}}{\sim} N(1, \sigma_x^2)$. By drawing the $x_{i,t}$ only once, we are able to simulate the case of fixed regressors, a standard assumption. For all the experiments we set $\sigma_x^2 = 0.01$; this value was chosen so that given our $\beta = 5$ and variance on the error terms, $\sigma_\varepsilon^2 = 1$, the average t statistic on the unit by unit OLS estimates would be approximately 2 when $T = 20$.

Then on each replication of the simulation, we drew the N unit parameters independently from a normal distribution with a set mean and variance, so $\beta_i \overset{\text{ind}}{\sim} N(\beta, \gamma^2)$; later experiments examined different ways of drawing the β_i. Since there is only a single covariate in our simulations, $\boldsymbol{\Gamma}$ is scalar; for interpretive reasons it is simplest to refer to the variance of the random coefficients as γ^2. For all of the experiments β was fixed at 5. We then generated the $y_{i,t} = \beta_i x_{i,t} + \varepsilon_{i,t}$ where $\varepsilon_{i,t} \overset{\text{ind}}{\sim} N(0, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon^2 = 1$.

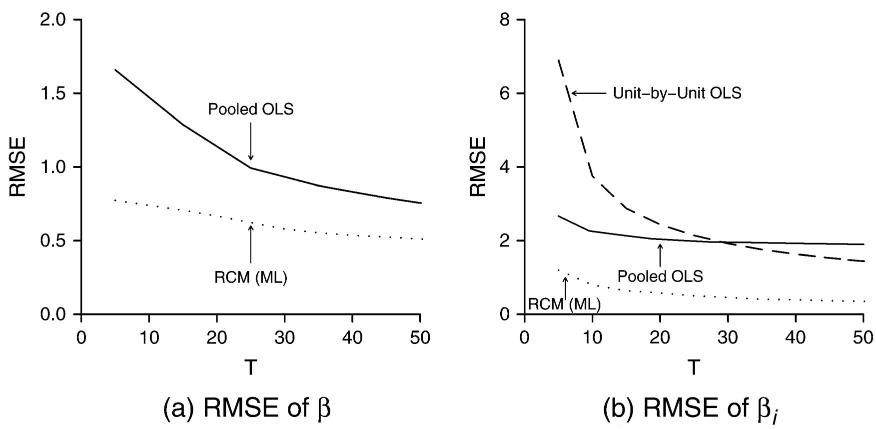

Fig. 1 Comparison of RMSE for RCM and OLS estimators of β and β_i as T varies from 5 to 50. For all runs of the experiment, $N = 20$, $\beta = 5$, $\gamma = 1.8$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$.

Unlike our previous work, we do not focus on the accuracy of the standard errors. This is because if the data are generated by an RCM, then we know from theory that the RCM standard errors are more accurate than the OLS standard errors. Given the T's we use, a comparison of the unit by unit standard errors is of little interest. In previous work, there was a trade-off of accuracy of standard errors and efficiency of estimation. As we shall see, this is not the case for the RCM standard errors, and hence we present no simulation results on the accuracy of standard errors.

4 Monte Carlo Results

Our first results compare the RCM (maximum likelihood) and OLS estimates of the overall β in equation (3) when the data are generated by an RCM (with the standard deviation of the β_i, $\gamma = 1.8$). Figure 1a shows the RMSE of the estimate of β as a function of T.

The results are clear. The RCM estimate of β always is more accurate than the corresponding OLS estimate, and the gain from using the RCM estimate is nontrivial. The advantage of the RCM is greatest for the smallest T; at our minimal $T = 5$, the RCM is twice as accurate as is OLS. But even for our largest $T = 50$, the RCM is still 25% more accurate than OLS. Thus, the efficiency gain of the RCM over OLS for estimating β is substantial and makes it worthwhile to use the RCM to estimate the overall β if we think there is any variation in the unit β_i.

The next experiment asks if we can be misled by the RCM if indeed there is no variation in the unit β_i. If $\gamma = 0$ pooled OLS must dominate the RCM (since the OLS model is correct and has one fewer parameter to estimate). But even in this extreme situation, the efficiency advantage of OLS is trivial. For example, when $T = 20$, the RMSE for the pooled OLS of β is 0.4858; for the RCM it is 0.4862. With even a tiny γ the RCM becomes superior to pooled OLS and that superiority of course grows with γ. But it is good to know that even when the pooled model is correct, the RCM performs at least 99% as well as pooled OLS. There is no fear that the RCM can mislead analysts into finding parameter variation when there is none.

Turning to the estimates of the unit β_i, we have three different estimators: the unit by unit OLS estimates, the RCM estimates, and the pooled OLS estimates (so each β_i is

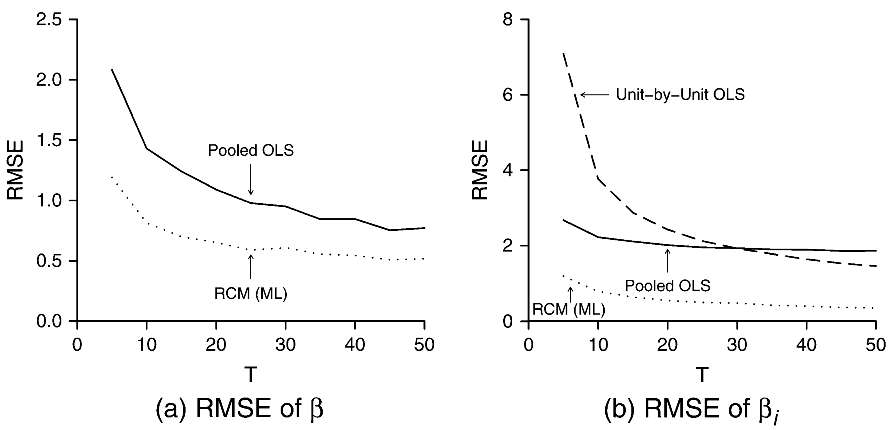

Fig. 2 Comparison of RMSE for RCM and OLS estimators of β and β_i as T varies from 5 to 50. For all runs of the experiment, $N = 20$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$. The β_i were drawn as $\Gamma(5, 1.8^2)$.

assumed to equal the common estimate of β). Keeping all parameters as in Fig. 1a, in Fig. 1b we show the RMSE of the three estimators of the β_i.

Until T gets large (about 30), the pooled estimator does a better job of estimating the individual β_i's than does unit by unit OLS; Although eventually unit by unit OLS must give better estimates than the pooled estimator, there are many practical situations where investigators would prefer the pooled to the unit by unit estimates of the β_i. Even for $T > 30$, the gains of the unit by unit estimator are not large (about 10%).

But we need not detain ourselves on this comparison since the RCM estimate of the individual β_i's is superior to either the fully pooled or unit by unit estimates. For all T's the gain from using the RCM is substantial, with that gain (over the better of the two OLS-based estimates) being over 50%. If one cares about the estimates of the individual β_i, these experiments show that one should use the RCM and not OLS.

We have, so far, presented experiments where the RCM is likely to fare the best (though we have shown that even if there is no unit to unit variation, the RCM does not mislead). But what happens if there is unit to unit parameter heterogeneity that is not normal? We first examine what happens if the β_i are drawn from a distribution that both is asymmetric and has a longer tail than the normal. Thus, we simulated the draws of the β_i from a gamma distribution. We chose the shape and scale parameters of the distribution to match our previous experiments, so the mean and standard deviation of the gamma distribution were set to be 5 and 1.8, respectively (the gamma distribution parameters were set to 7.8126 and 0.64, respectively); other parameters were identical to those in the normal experiment. In Fig. 2, we compare OLS and RCM estimates of both β and β_i.

The results are almost identical to those seen in Fig. 1 where the β_i were normally distributed. The superiority of the RCM, both for estimating β and β_i, is as strong when the β_i are generated by a distribution that is highly nonnormal as when the β_i are generated in a manner that exactly corresponds to the assumptions of the RCM. Although obviously one set of experiments cannot provide conclusive evidence, it surely does not appear that the assumption that the β_i are distributed normally is a critical assumption.

We now consider a more extreme situation. Suppose that one or two of the unit β_i are outliers. Can the RCM pick those up better than unit by unit OLS? In Fig. 3, we consider estimating the β_i in a situation where 18 of the unit β_i's are 5 and 2 are outliers, ranging from 5 to 10. Can the RCM do a better job of picking up those outliers than OLS?

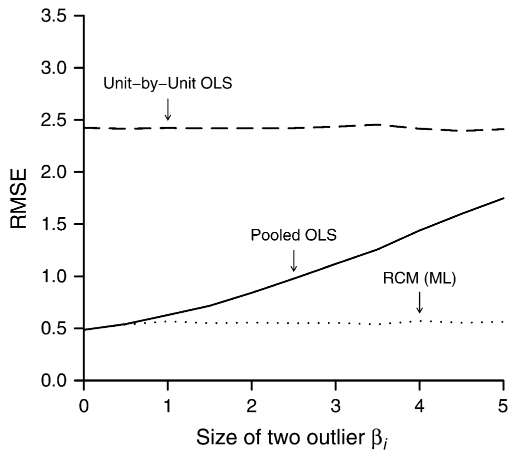

Fig. 3 Comparison of RMSE for RCM and OLS estimators of β_i where the last 18 $\beta_i = 5$ and the first two vary from 5 to 10. For all runs of the experiment, $N = 20$, $T = 20$, $\sigma^2_\varepsilon = 1$, and $\sigma^2_x = 0.01$.

(We do not consider estimating β here since we are not interested in an average of the 18 nonoutliers and the 2 outliers.)

The RCM continues to perform well in this situation. As we induce two bigger and bigger outliers in the β_i, the RMSE of the RCM estimates of the β_i do not change. The highly nonnormal way we drew the 20 β_i caused no problems for the RCM estimation of the β_i. Thus, at least for this case, even when the β_i do not look at all like equation (3), the RCM still yields good estimates of the β_i. This is reassuring since the normality of the β_i is clearly an assumption driven more by mathematical convenience than by empirical reality. This also means that the RCM provides a method for assessing whether some units are outliers (or, more generally, whether the β_i can be seen as draws from a common distribution). Obviously, pooled OLS is not suitable for this task. Although for large enough T unit by unit OLS would work for this purpose, the RCM is clearly superior for finding outlying units given the T's we see in TSCS data.

5 Conclusion

There is little doubt, as Western noted, that the RCM should appeal to scholars of comparative political economy (broadly defined) who are not so naive as to assume that all countries (or units) are identical but who are sufficiently committed to comparative analysis that they cannot assume that all units are unique. The RCM appears to offer the analyst a flexible middle position, one that allows the data to tell us how heterogeneous the units are.

Our Monte Carlo experiments indicate that the RCM performs very well, both for the estimation of the overall $\boldsymbol{\beta}$ and the unit-specific $\boldsymbol{\beta}_i$. This assumes that estimation is by some variant of maximum likelihood. As we see in the appendix, the performance of the FGLS RCM estimator is quite poor. We note that there are a number of different ways to maximize the likelihood; since we have investigated only one such way, we cannot compare the various alternatives. We would assume the estimates are similar, so those who like Markov chain Monte Carlo or simulated likelihood are probably quite safe in using those techniques to explore the likelihood.

Not only does the RCM perform well in terms of mean squared error but our results also indicate that it does not mislead, in the sense that if there is little or no parameter heterogeneity, the RCM will not falsely find it. There are no places where simpler methods (nontrivially) dominates the RCM in terms of mean squared error. Thus, unlike other complicated routines, there seems little danger from starting with the RCM. It also appears that the RCM handles parameter dispersion that is highly nonnormal, even though the model assumes normality. How robust the estimators are to all the possible violations of assumptions is well beyond the scope of any one article, but none of our experiments (both reported and unreported) lead us to expect the RCM to be a brittle estimator.

If the RCM indicates little parameter heterogeneity (that is, a small and/or insignificant γ), then we might wish to go back to simpler, OLS-based methods since these will be more flexible. Thus, analysts could use a pretest strategy, running the RCM, testing the null hypothesis that $\gamma = 0$ (or some or all of the diagonal terms of $\boldsymbol{\Gamma}$ are zero), with a subsequent reanalysis by OLS if that null is not rejected (or in the multivariate case the assignment of some coefficients to the fixed category). But the R implementation of the RCM is quite flexible; it can handle a variety of correlated error structures and it extends very nicely to nonlinear models.[7] But if, for example, spatial issues are of interest and there is not much parameter heterogeneity, then it might be the case that starting with a simple linear model would allow researchers to better deal with spatial issues. Such trade-offs will always exist for any method.

So why do applied TSCS analysts in political science seem not to use the RCM? One reason is that the biggest gains come from estimating the unit-specific parameters, and many comparativists may be primarily interested in the overall parameters (including general determinants of the variability as in equation 7). But it is hard to believe that there is not some interest in the unit-specific parameters, even if only to check for outlying units, and the gains from using the RCM to estimate the overall $\boldsymbol{\beta}$ are nontrivial.

It also may be the case that TSCS analysts used the FGLS methods of Swamy and Hsiao; this may have been done because they have been long implemented in packages commonly used by political scientists. Such analysts might have then thought the RCM to be of little value because of the misleading estimates of the FGLS routines. For such analysts, we strongly recommend that the RCM, estimated via some variant of maximum likelihood, be reconsidered.

The RCM is not a panacea. Although it allows researchers to check for homogeneity across units, it is not the only way to do so. Thus, we have recommended in other articles (Beck 2001) that analysts assess homogeneity via cross-validation (leaving out a unit at a time); this still makes sense. The R routines come with a large variety of diagnostics to assess whether the assumptions of the RCM hold; these should be rigorously used. Clearly, residuals need to be checked for time-series and spatial issues.

In this article, we have looked only at the simplest case, one independent variable with a random coefficient and a single, fixed constant term. We do not know how the RCM will perform in more complicated situations. What happens if one has 10 independent variables or random effects? It is almost certainly true that "nuisance" control variables should probably be treated as fixed. And since the full $\boldsymbol{\Gamma}$ covariance matrix has approximately $K^2/2$ parameters to estimate, it is likely to be problematic to estimate a full $\boldsymbol{\Gamma}$ matrix for 10 independent variables. But one can specify simpler structures, particularly by assuming that the random coefficients are all independent (so that only the diagonal elements of $\boldsymbol{\Gamma}$

[7]Moreover, because it is open source, it is constantly being extended. Thus, for example, while this article was being revised a new S package was introduced that added smoothing splines to the nonlinear mixed estimation routines; Bates and colleagues have also recently released a new function, lmer (in their matrix package), which extends the lme function to generalized linear models and improves numerical performance.

need to be estimated). Although one may carp at that assumption, current practice consists of making the much stronger assumption of homogeneous coefficients. Just because the RCM allows for a huge amount of flexibility, it does not mean that we need to use all that flexibility; general experience shows that imposing some structure on a flexible model often improves matters. But specific recommendations must await future work.

It is likely that the RCM will be more advantaged over the OLS methods as we move to more complicated situations. The RCM shows greatest advantage for the smaller values of T. But adding more independent variables is similar to decreasing the effective sample size. Similarly, serial correlation of the errors can be seen as making the effective sample size considerably smaller than the nominal sample size of T (for each unit). Thus, in complicated situations, we speculate that the RCM, suitably constrained, will be even more valuable than in the simple situations we have considered here.

The overall conclusion from our experiments is clear. Recent work has made full maximum likelihood estimation of the RCM both feasible and available to researchers using standard packages. These methods perform well, when the RCM model fits the data generation process, and do not perform poorly and do not mislead, when either the parameters do not vary across units or vary in some ways that are quite different from the assumptions underlying the RCM. TSCS analysts should think seriously about including the RCM, estimated via maximum likelihood, as one of their principal empirical modeling tools. The costs are low and the gains can be high. Perhaps the best analogue here is the heteroskedasticity consistent standard errors: when not needed they do no harm and when needed they do a lot of good. So although the RCM should never be arbitrarily chosen as a stopping point, it is almost always a good starting point for TSCS analysts.

Appendix A: The Poor Finite Sample Properties of the Swamy-Hsiao FGLS Estimator

As noted in the text, a standard RCM estimator (e.g., Hsiao 2003, 144–51) is a particular implementation of Swamy's (1971) FGLS estimator. Although we have no doubt that analysts should use maximum likelihood (or related) estimators for the RCM model, the FGLS method is still used, is implemented in standard statistical packages, and is considered viable in standard texts (as shown in our discussion in Section 2 of Hsiao's leading text). Since the poor finite sample properties of the FGLS estimator are clear and we have laid out the details in an unpublished paper (Beck and Katz 2004), here we only lay out the brief argument and present key Monte Carlo results. Since the derivation of the FGLS estimator is in Hsiao (2003, 145–6), we only present the final results, referring the reader to Hsiao for the derivations.

To understand the issues, recall that the estimate of the overall mean from the pooled OLS is inefficient because it does not use all of the information in the structure of the model. An FGLS estimate of $\boldsymbol{\beta}$ builds on the RCM as a linear model with a complicated error structure, as in equation (6). As with all FGLS estimators, we start with the consistent but inefficient estimator of $\boldsymbol{\beta}$, OLS. We then estimate the parameters of the nonspherical variance-covariance matrix of the errors and then use this estimate in the standard generalized least squares transformation. Hsiao shows that everything can be built up from the OLS estimators, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_i$ (OLS estimates are designated by carets with tildes for FGLS).

Hsiao shows that the FGLS estimator can be seen as a weighted sum of unit by unit OLS estimators of the $\boldsymbol{\beta}_i$,

$$\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{W}_i \hat{\boldsymbol{\beta}}_i, \tag{A1}$$

$$W_i = \left\{ \sum_{i=1}^{N} [\mathbf{\Gamma} + \mathbf{V}_i]^{-1} \right\}^{-1} [\mathbf{\Gamma} + \mathbf{V}_i]^{-1}, \tag{A2}$$

$$\mathbf{V}_i = \sigma_i^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1}, \tag{A3}$$

where $\mathbf{X}_i$ denotes the matrix of observations for unit i, $\mathbf{\Gamma}$ is the variance matrix of random coefficients (equation 3), and $\mathbf{V}_i$ is the variance of $\hat{\boldsymbol{\beta}}_i$.

Since this is FGLS, we do not know $\mathbf{\Gamma}$ or σ_i^2, but instead we use estimates of them from the initial OLS regressions. Swamy and Hsiao suggest the usual OLS estimate of σ_i^2. The question is how to estimate $\mathbf{\Gamma}$? If we could directly observe the β_i, we could use the N draws to construct an estimate of the covariance matrix in the usual fashion:

$$\widetilde{\Gamma} = \frac{1}{N-1} \left(\sum_{i=1}^{N} \left(\boldsymbol{\beta}_i - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\beta}_i \right)' \left(\boldsymbol{\beta}_i - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\beta}_i \right) \right). \tag{A4}$$

Since averaging is done over N, estimation will improve as N grows; this is very important for panel analysts but of cold comfort for TSCS analysts.

The problem is we do not observe $\boldsymbol{\beta}_i$; we have, instead, only noisy estimates, $\hat{\boldsymbol{\beta}}_i$. So although we might consider just substituting $\hat{\boldsymbol{\beta}}_i$ for $\boldsymbol{\beta}_i$ in the definition of $\tilde{\mathbf{\Gamma}}$, this would lead to an overestimate of the amount of variation in $\boldsymbol{\beta}_i$ since much of the variation in the $\hat{\boldsymbol{\beta}}_i$'s is caused not by "real" parameter variability but purely by sampling error. We can correct for this sampling variability by subtracting it off. Swamy thus suggested that a plausible estimator of $\mathbf{\Gamma}$ to use in the FGLS estimation is

$$\tilde{\mathbf{\Gamma}} = \frac{1}{N-1} \left(\sum_{i=1}^{N} \left(\hat{\boldsymbol{\beta}}_i - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_i \right)' \left(\hat{\boldsymbol{\beta}}_i - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_i \right) \right) - \frac{1}{N} \sum_{i=1}^{N} \hat{\sigma}_i^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1}. \tag{A5}$$

There is, however, a problem with this estimator: in finite samples $\tilde{\mathbf{\Gamma}}$ need not be positive definite, a necessary requirement for it to be a well-defined covariance matrix. This is because we are subtracting off the estimated sampling variance that can be large for the typical values of T found in TSCS data.

The question is how to ensure that $\tilde{\mathbf{\Gamma}}$ is positive definite? Hsiao (2003, 146), building on Swamy, suggests that the second term of $\tilde{\mathbf{\Gamma}}$ be dropped. The rationale for this is asymptotic, that is, as NT gets large the sampling variance goes to zero. This fix is not correct in finite samples as it will tend to overestimate $\mathbf{\Gamma}$. The question for TSCS analysts is how badly does this omission of sampling variance affect the estimate of $\mathbf{\Gamma}$ and does this cause any problems in the estimates of $\boldsymbol{\beta}_i$ for typical values of N and T seen in actual research situations? Since these are problems in finite samples, we will have to assess the claims via Monte Carlo simulations.

We replicate our standard experiment for assessing the estimate of the β_i but this time simply comparing the Swamy-Hsiao FGLS estimates to the maximum likelihood estimates (replicating the conditions of Fig. 1b). Results are in Fig. 4a.

Clearly the FGLS estimator performs considerably worse than the maximum likelihood estimator. As T gets larger this inferiority declines, but even for our largest $T = 50$, the FGLS RMSE is twice that of the maximum likelihood estimator.

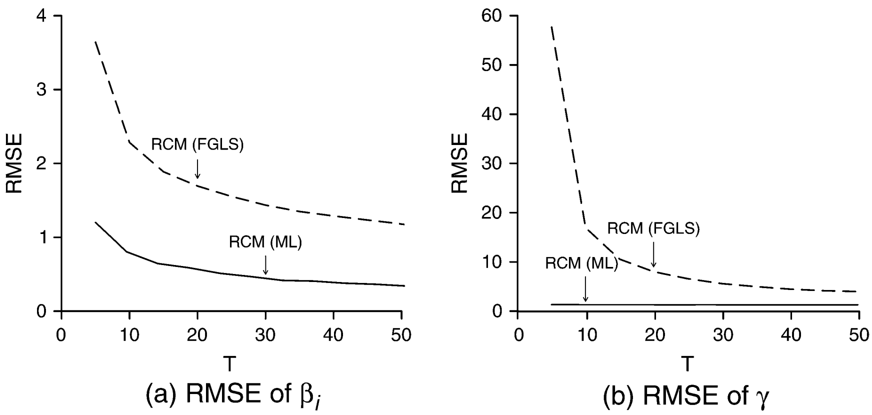

Fig. 4 Comparison of RMSE for RCM and FGLS estimators of β_i and γ as T varies from 5 to 50. For all runs of the experiment, $N = 20$, $\beta = 5$, $\gamma = 1.8$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$.

Why does the FGLS estimator perform so poorly? As noted above, it is because in finite samples the variance term is underestimated. In Fig. 4b we compare the RMSE of the estimate of γ for the FGLS and maximum likelihood procedures. For very small $T \leq 10$ the FGLS estimate of γ is horrible; even past that, the RMSE for the FGLS estimator is never less than five times as large as for the maximum likelihood estimator. The Hsiao assumption that the sampling variance is zero, used to ensure positive definiteness of a variance-covariance matrix, is the culprit here.

When the FGLS estimators for the RCM were developed, there were no computationally feasible alternatives to FGLS. The world has changed, and maximum likelihood is now quite feasible. The maximum likelihood estimator strongly dominates the FGLS estimator (and always will be better in all possible ways). Given that the maximum likelihood estimator is now implemented in commonly used packages, there is simply no reason for anyone to again consider the Swamy-Hsiao FGLS estimator. It may once have been the best available compromise, but its compromise is both costly and completely unnecessary today.

References

Adolph, Christopher, Daniel M. Butler, and Sven E. Wilson. 2005. Which time-series cross-section estimator should I use now? Guidance from Monte Carlo experiments. Paper presented at the 2005 annual meeting of the American Political Science Association, Washington, DC.

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. 2004. Random coefficient models for time-series–cross-section data. Social Science Working Paper 1205, Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, September, 2004.

Efron, Bradley, and Carl Morris. 1975. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* 70:311–9.

Garrett, Geoffrey. 1998. *Partisan politics in the global economy.* New York: Cambridge University Press.

Greene, William. 2003. *Econometric analysis.* 5th ed. Upper Saddle River, NJ: Prentice Hall.

Hsiao, Cheng. 2003. *Analysis of panel data.* 2nd ed. New York: Cambridge University Press.

Kedar, Orit, and Phillip Shively. 2005. Introduction to the special issue. *Political Analysis* 13:1–4.

Pinheiro, José C., and Douglas M. Bates. 2000. *Mixed effects models in S and S-Plus.* New York: Springer.

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

Swamy, P. A. V. B. 1971. *Statistical inference in random coefficient models*. New York: Springer-Verlag.

Wallerstein, Michael, and Bruce Western. 2000. Unions in decline? What has changed and why. *Annual Review of Political Science* 3:355–77.

Western, Bruce. 1997. *Between class and market: Postwar unionization in capitalist democracies*. Princeton, NJ: Princeton University Press.

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