Advance Access publication April 10, 2009

Political Analysis (2009) 17:177–190
doi:10.1093/pan/mpp004

Modeling Certainty with Clustered Data: A Comparison of Methods

Kevin Arceneaux
Department of Political Science, Institute for Public Affairs, Faculty Affiliate, Temple University, 453 Gladfelter Hall, 1115 West Berks Street, Philadelphia, PA 19122
e-mail: kevin.arceneaux@temple.edu (corresponding author)

David W. Nickerson
Department of Political Science, University of Notre Dame, 217 O'Shaughnessy Hall, Notre Dame, IN 46556

Political scientists often analyze data in which the observational units are clustered into politically or socially meaningful groups with an interest in estimating the effects that group-level factors have on individual-level behavior. Even in the presence of low levels of intracluster correlation, it is well known among statisticians that ignoring the clustered nature of such data overstates the precision estimates for group-level effects. Although a number of methods that account for clustering are available, their precision estimates are poorly understood, making it difficult for researchers to choose among approaches. In this paper, we explicate and compare commonly used methods (clustered robust standard errors (SEs), random effects, hierarchical linear model, and aggregated ordinary least squares) of estimating the SEs for group-level effects. We demonstrate analytically and with the help of empirical examples that under ideal conditions there is no meaningful difference in the SEs generated by these methods. We conclude with advice on the ways in which analysts can increase the efficiency of clustered designs.

1 Introduction

Researchers in political science are often confronted by data in which the units of analysis are observationally grouped. Citizens are grouped by neighborhoods. State supreme court judges are grouped by the 50 separate state courts on which they serve. Presidential statements are grouped by the individual presidents who made them. In each of these instances, it is plausible that two observations within a group may be more similar to each other than to another observation in a different group. Analysts should take heed when variance in the outcomes of interest can be explained by grouping. For instance, if neighbors tend to vote in similar ways or if President Clinton's statements are similar with each other but not with statements made by President Carter, analysts risk severely underestimating the uncertainty attached to their causal estimates by ignoring the clustered architecture of their data.

Authors' note: We would like to thank Bob Erikson, Chris Zorn, and the anonymous reviewers for helpful comments and suggestions. We would also like to thank Robert Brown and Lawrence Broz for generously sharing data. All errors remain our own.

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

The standard approach to estimating the variance-covariance matrix of the data assumes one value characterizes the variance across observations (i.e., homoskedastic variance) and that no observations are correlated with other observations (i.e., observations are independently distributed). However, if observational units are correlated within well-defined clusters and those clusters differ from other clusters in meaningful ways, both the homoskedastic and independence assumptions are violated. As we formally demonstrate in the next section, when units are positively correlated within clusters (a typical case in political science), invoking these standard assumptions causes researchers to underestimate the standard errors (SEs) of causal estimates even in the presence of low levels of intracluster correlation. From an epistemological point of view, this is problematic because it increases the probability of committing a Type I error.

Although survey researchers have become more cognizant about the need to adjust SEs to model surveying sampling techniques (Kish 1965; Stoker and Bowers 2002), the issue of clustering is often given insufficient attention across other areas of research (for useful exceptions, see Arceneaux 2005; Zorn 2006; Green and Vavreck 2008). In this paper, we compare four practical approaches that researchers can choose to produce a more accurate estimate of their SEs: clustered robust estimation, random effects, hierarchical modeling, and aggregation to the level of clustering. Although each approach is familiar to political methodologists, little guidance is available to help researchers choose among these methods. More typically, scholars champion a preferred methodology and trumpet its virtues. Furthermore, the extant discussion about these estimators focuses upon point estimation and largely ignores how uncertainty is modeled.

We fill this void and offer practical advice in selecting models to accommodate clustered data. To illustrate when and how one analyzing clustered data should adjust SEs, we begin with a mathematical explanation of the problem and analytically derive a number of lessons about the effects that clustering has on the estimation on model uncertainty. We then demonstrate some of these key lessons with the help of empirical examples. Intuitively, one would expect that analyzing clustered data at the individual level with appropriate methods (i.e., clustered SEs, random effects, or hierarchical linear model [HLM]) would increase the precision of estimates relative to aggregating to the cluster level. However, we find that this is not always the case and derive the conditions under which it is. Clustering, random effects, hierarchical, and aggregation models arrive at very similar SEs. If researchers are only interested in estimating group-level effects and their data meet the strong assumptions underlying these approaches, there is little reason to select one method over another. We conclude with a discussion of ways in which analysts can increase the efficiency of their model estimates through good design practices and outline special cases in which scholars should consider adopting one method over another.

2 Options in Analysis

2.1 *Basic Problem*

Assume there are N individuals clustered into G groups, $1 < g < G$, where each group comprises n_g individuals so that $\sum_{g=1}^{G} \sum_{i=1}^{n_g} i = \sum_{g=1}^{G} n_g = N$. To relieve stylistic monotony, we will use the terms *groups* and *clusters* interchangeably. We are interested in estimating the effect, β, of a particular variable of theoretical interest, denoted T, on an outcome variable, Y. For the sake of exposition and without loss of generality, suppose T is dichotomous (i.e., $T = 1$ when the variable of theoretical interest is present and $T = 0$ when the variable of theoretical interest is absent). Let each g be assigned (either by the researcher or by the world) to different values of T, such that all individuals in a given group have the same value on T. For the purpose of estimating β, consider the basic model

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

$$Y_{ig} = \alpha + \beta T_g + \epsilon_{ig},$$

where α is the intercept or average level of the outcome variable when $T = 0$ and ϵ_{ig} are the idiosyncratic causes of the outcome variable for each individual. Furthermore, assume that $\text{cov}\left(T_g, \epsilon_{ig}\right) = 0$. This assumption would be met if the data were derived from an experiment in which the researcher randomly assigned groups to values of the variable of theoretical interest or, in the case of an observational study, if the researcher were able to account for all relevant covariates that are correlated with both T and Y. This standard assumption underlies all empirical approaches to estimating causal effects, and if it is met, ordinary least squares (OLS) yield unbiased point estimates for β. Much has been written on the epistemological questions surrounding causal inference. We wish to bracket those questions here and, instead, focus on approaches to estimating the degree of uncertainty surrounding the estimate of $\hat{\beta}$. It is not immediately apparent whether the unit of analysis should be the unit of data collection, i, or the unit of assignment, g.

Conducting the analysis on the individual level assumes N independent observations. The variance associated with an individual OLS estimate of β is

$$\text{Var}\left(\beta_{\text{OLS}}\right) = \frac{\sigma^2}{\sum_g \sum_i \left(T_{ig} - \bar{T}\right)^2}, \tag{2}$$

where $\sigma^2 = \sum_g \sum_i \epsilon_{ig}^2$.[1] However, this formula for the variance assumes the errors to be independent, which we know to be false from the construction of the data. In fact, $T_{gi} - T_{hj} = 0 \; \forall \; g = h, \; i \neq j$. Thus, a naive individual-level analysis using OLS will overstate the certainty of our estimates.

A more conservative approach is to aggregate up to the unit of assignment by taking the mean for each group (see equation 3).

$$\bar{Y}_g = \alpha + \beta \bar{T}_g + \bar{\epsilon}_g. \tag{3}$$

The aggregate analysis is well behaved with $E\left(\bar{\epsilon}_g\right) = 0$ and $\text{var}\left(\bar{\epsilon}_g\right) = \frac{\sigma^2}{n_g}$ (Kmenta 1997, 368). The variance for OLS estimates of β on the aggregated data is calculated by

$$\text{var}\left(\beta_{\text{aggregated}}\right) = \frac{\sigma^2}{\sum_g n_g \left(\bar{T}_g - \bar{T}\right)^2}. \tag{4}$$

After a few algebraic manipulations, the variance for the aggregated estimate can be compared to the variance of the naive individual-level regression analysis:

$$\frac{\text{var}\left(\beta_{\text{aggregated}}\right)}{\text{var}(\beta_{\text{OLS}})} = 1 + \frac{\sum_g \sum_i \left(T_{ig} - \bar{T}_g\right)^2}{\sum_g n_g \left(\bar{T}_g - \bar{T}\right)^2}. \tag{5}$$

The numerator of the second term on the right-hand-side of equation (5), $\sum_g \sum_i \left(T_{ig} - \bar{T}_g\right)^2$, represents the variance of terms within clusters. The denominator of the second term on the right hand of equation (5), $\sum_g n_g \left(\bar{T}_g - \bar{T}\right)^2$, is the weighted variance of means across clusters. Equation (5) nicely illustrates two instructive points.

[1] Typically, $\sigma^2 = \frac{1}{N} \sum_i \sum_g \epsilon_{ig}^2$. However, the denominator is the variance of T, which is typically estimated $\text{var}\left(T\right) = \frac{1}{N} \sum_i \sum_g \left(T_{ig} - \bar{T}\right)^2$. For notational ease, the $\frac{1}{N}$ is omitted from both numerator and denominator.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

The first lesson is that in most cases where data are clustered into groups, a naive individual-level OLS analysis will underestimate the true variance of the estimate. Both the numerator and the denominator of the second term on the left-hand-side are positive, so the variance will generally be smaller than it would be if the data were aggregated up to the level of assignment.

The second lesson is that if the majority of the variance in the subject population is across groups (i.e., $\sum_g n_g\left(\bar{T}_g-\bar{T}\right)^2$ is large), then there is little reason to examine the individual-level data. Conversely, in instances where there is little variance within groups (i.e., $\sum_g \sum_i \left(T_{ig}-\bar{T}_g\right)^2$ is small), there is no gain in efficiency from individual-level analysis. In the current framework, where clusters are assigned to values of T, there is no variance within clusters and, thus, there is no efficiency gain whatsoever from individual-level analysis because $\sum_g \sum_i \left(T_{ig}-\bar{T}_g\right)^2 = 0$. Adding covariates to the analysis or varying the value of T within clusters would alter this result.

The flipside of the second lesson is that when there is a great deal of variance within each cluster (i.e., $\sum_g \sum_i \left(T_{ig}-\bar{T}_g\right)^2$ is large), there would appear to be efficiency gains from moving from the aggregate to the individual level (we derive the precise conditions under which this is so below). The next three sections describe strategies for individual-level analysis that appropriately adjust the variance estimates to account for the clustered nature of the treatment application.

2.2 *Clustered SEs*

The first method makes no changes to the estimation procedure from OLS but adjusts the SEs to account for correlation between individuals who are nested within clusters. The first step in the process is to decompose the residual error term, ϵ_{ig}, into the part due to the cluster, γ_g, and the part due to the individual, u_{ig}, both with mean zero. Using the variance notation developed in the last section, $E(\epsilon_{ig}^2) = \sigma^2 = \sigma_\gamma^2 + \sigma_u^2$. By assumption, the group component is assumed to be uncorrelated across groups (i.e., $E\left(\gamma_g\gamma_h\right) = 0 \,\forall\, g \neq h$). The individual-level error component is assumed to be independent both within and across cluster. That is, $E\left(\epsilon_{ig}\epsilon_{jh}\right) = 0 \,\forall\, g, h \; i \neq j$. Thus, the model presented in equation (1) can be rewritten as

$$Y_{ig} = \alpha + \beta T_g + \gamma_g + u_{ig}. \tag{6}$$

In principle, the bias from clustered data is a matter of unmodeled group-level error. If one could collect variables that perfectly captured the group-level error, then the SEs for the treatment of interest using naive OLS would be accurate. In practice, one can never be certain that all the potential causes of group-level differences have been accounted for and the techniques that allow for group-level errors should be utilized. As a research design principle, however, the improvement in statistical efficiency from collecting informative covariates can be equal to adding new observations or clusters to the analysis and should not be overlooked.

The critical concept when analyzing clustered data is the ratio of variance within clusters to overall variance in the model. This concept is typically referred to as the intracluster correlation coefficient (ICC),[2] ρ, where $\rho = \frac{\sigma_\gamma^2}{\sigma_\gamma^2+\sigma_u^2}$. If most of the variance is found between individuals and cluster means do not vary much (i.e., σ_γ^2 is small), then ρ will be near zero. Intuitively, a small intracluster correlation implies that the clusters explain little and that

[2]The intracluster correlation coefficient is also referred to as the intraclass correlation coefficient.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

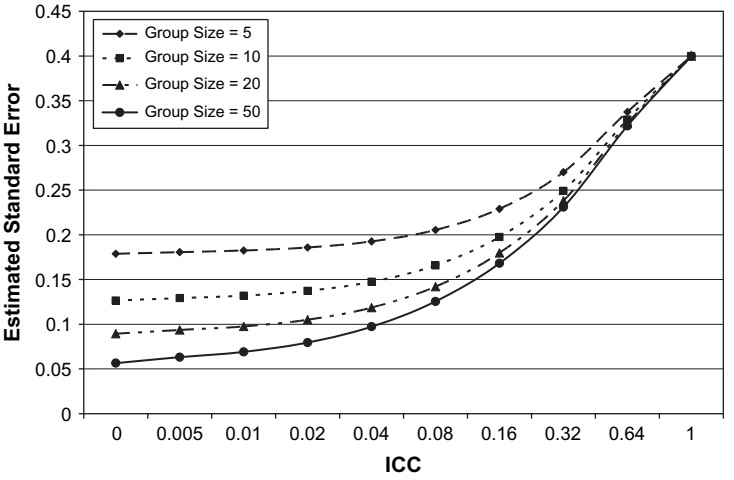


Fig. 1 The effect of intracluster correlation on statistical efficiency.

the effective sample size is closer to the number of individuals, *N*, than the number of groups, *G*. In such instances, analysis at the individual level may gain efficiency. In contrast, if individuals are relatively homogenous within clusters and each cluster is markedly different (i.e., σ_{γ}^2 is large), then ρ will also be large and the effective sample size is closer to *G* than *N*. High intracluster correlation means the individual-level OLS estimates will be severely biased and in need of adjustment. The size of the adjustment depends upon the ICC and the number of individuals in each group.[3] The naive variance formula (equation 2) can easily be adjusted by including a ''variance inflation factor'' as presented in equation (7) (see Kish 1965; Donner 1998, 98; or Murray 1998, 362).[4]

$$\mathrm{Var}\left(\beta_{\mathrm{Clustered}}\right) = \frac{\sigma^2\left\{1+\left(\bar{n}_g-1\right)\rho\right\}}{\sum_g \sum_i \left(T_{ig}-\bar{T}\right)^2}. \tag{7}$$

As defined above, the variance inflation factor is always positive.[5] That is, $1+\left(\bar{n}_g-1\right)\rho \geq 0$. With this restriction in place, the first lesson to take from equation (7) is that naive OLS (equation 2) underestimates the degree of uncertainty surrounding estimates of the treatment effect. As ρ increases, the OLS variance estimates are biased to a greater extent (see Fig. 1). Even in the presence of low levels of intracluster correlation and moderate group sizes, SEs can change dramatically. When the intracluster correlation changes from zero to

[3]To simplify the presentation, equation (7) assumes groups to be of equal size. To account for clusters of different sizes, iterative processes can be used to sum the each one of the clusters.

[4]The term $1+\left(\bar{n}_g-1\right)\rho$ was initially termed the ''design effect'' by Kish (1965), but ''variance inflation factor'' has gained popularity in the medical sciences (e.g., Donner 1998).

[5]The scalar notation developed in the formula conceals the possibility of negative intracluster correlation. We adopted scalar notation because it is clearer and more accessible. The potential for negative intracluster correlation is more obvious using matrix algebra (e.g., Greene 2000, 567–72). Negative intracluster correlation can occur when there is little between-group variation in means but considerable unexplained variance within clusters. On the off chance that the intracluster correlation is negative, researchers will find that statistical adjustments for cluster produce *smaller* SEs than naive approaches do. In these rare cases, we recommend that analysts report the more conservative, larger SE estimates. Our first empirical example encounters this situation.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

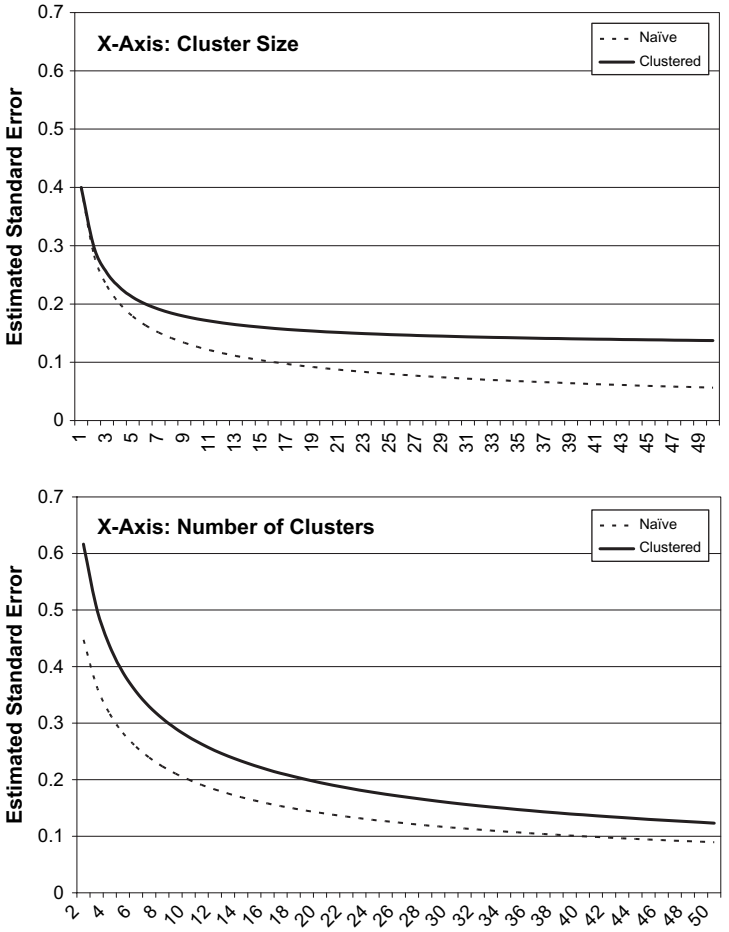


Fig. 2 Efficiency from increasing cluster size and the number of clusters.

0.04 with an average group size of 20, the SE inflates from 9 to 12 percentage points—a 33% increase.

Equation (7) also illustrates why adding individuals to each cluster (i.e., increase $\bar{n}_g$) matters less than adding clusters, (i.e., increase G) to the study. Namely, the size of the clusters appears in both the numerator and the denominator, whereas the number of clusters appears only in the denominator (see Fig. 2). The top panel of Fig. 2 depicts the gains in efficiency as the average number of individuals in each cluster increases with an intracluster correlation of 0.1. Moving from one to six individuals in each group increases the precision by 50%, but once each group has 10 individuals, the gains in efficiency are limited. In contrast, the gains in efficiency increase much more consistently as the number of clusters increases (see Fig. 2, bottom panel). The gain in efficiency from moving from 20 to 30 clusters (20%) is the same as moving from 9 to 33 observations per cluster. Even when intracluster correlation is only moderate, researchers gain statistical power more cost effectively by increasing the number of clusters rather than the number of observations per group.

Furthermore, Donner and Klar (2000) demonstrate that clustered SEs provide overly optimistic SEs when the size of clusters is small. The typical rule of thumb cut-off provided by the medical literature is that 20 clusters are sufficient for reliable estimates. In most

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

settings, the number of clusters will be set and unchangeable for the researcher, but in those instances where the researcher has control over the design of the study, dividing the individual-level population into a larger number of clusters is a good design principle.

Comparing equation (7) to the formula for the variance of aggregated estimates of the treatment effect (equation 4), it is not immediately apparent which estimator is more efficient. The term $1+\left(\bar{n}_g-1\right)\rho$ inflates the numerator of the clustered variance; however, the variance of group means in the denominator of the aggregated formula should also be smaller inflating the variance of the aggregated treatment effect estimate. Assuming homogenous cluster sizes, algebraic manipulations of equations (4) and (7) yields the following proposition:

Proposition: Without informative covariates and positive intracluster correlation, analyzing individual-level data with clustered SEs is more statistically efficient than analyzing data aggregated by clusters if and only if the variance inflation factor is smaller than one plus the ratio of the variance of the variable of interest within clusters to the variance of the variable of interest across clusters. That is, $1+\left(\bar{n}_g-1\right)\rho < 1+\frac{\sum_g\sum_i\left(T_{ig}-\bar{T}_g\right)^2}{\sum_g n_g\left(\bar{T}_g-\bar{T}\right)^2}$. There is no efficiency to be gained when the variable of interest is constant within clusters. (Proof in appendix.)

Since both methods account for the clustered nature of the data, large differences in precision are not anticipated in most circumstances.

2.3 *Random Effects*

The model presented in equation (6) is structurally identical to the well-developed random effects model used in panel data settings. Typically randomly effects are used to estimate correlations within countries or people who are measured in multiple time periods. By substituting clusters for people and subjects within the cluster for time periods, random effects models can be used to estimate cluster randomized experiments.

The generalized least squares techniques used to estimate the variance of treatment effects in random effects setting use formulas differing from equation (7). However, since the clustered and random effects models are structurally identical, the variance components will also be identical; we do not anticipate the estimated variance for clustered and random effects models to differ in a meaningful way. Simulations (not shown) and our empirical examples confirm this intuition.

2.4 *Hierarchical Modeling*

On the surface, the multilevel modeling technique popularized by Raudenbush and Bryk (2002) appears markedly different from the causal model posited in equation (6) for the clustered and random effects models. HLMs capture the full complexity of multilevel data by explicitly modeling both the individual level and the clustered level (see Steenbergen and Jones 2002, for a helpful introduction). Equation (8) presents the group randomized trials in an HLM model:

$$Y_{ig} = \alpha + u_{ig}, \alpha = \mu + \beta T_g + \gamma_g. \tag{8}$$

The intercept, α, differs for each cluster and is a function of an overall mean tendency, μ; the effect of the variable of theoretical interest, βT; and idiosyncratic characteristics of each

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

cluster, γ_g. Thus, HLM allows stochastic variation to be modeled both at the individual and the cluster level. Substituting in the cluster-level information for alpha, the reduced form of the system of equations is found to be

$$Y_{ig} = \mu + \beta T_g + \gamma_g + u_{ig}. \quad (9)$$

Equation (9) is structurally identical to equation (6). Thus, all the variance components are the same for the hierarchical model as the clustered model, and we should not expect the variance of treatment effect estimates to differ.

Thus, although naive individual-level OLS results will consistently underestimate uncertainty in clustered data, clustered SEs, random effects, and hierarchical models, all adequately account for the structure of the data. In the remainder of the paper, we illustrate the key lessons drawn from our analytical treatment of the clustered data problem using empirical examples. In doing so, we hope to move beyond abstract statements about the properties of clustered SEs and demonstrate how clustered data affect estimates of model uncertainty in practice. Two disparate examples are showcased purposively to underscore the breadth of the problem. Clustering is not an issue that only bedevils scholars who study surveys and mass political behavior, it is one that touches on scholarship across the far flung regions of the field.

3 Empirical Examples

3.1 *State Party Organizations and Turnout*

In the past 25 years, over 100 articles have been written on the effects of U.S. state political characteristics on participation. There are a number of reasons to suspect that the voting behavior of individuals is more correlated within states than across. Voters in a state share the same political history, constellation of media markets, and set of statewide political elites—all of which are distinctive across states and difficult to model. Nevertheless, many scholars do not account for state-level clustering when analyzing the effects state political characteristics and, thus, many of the conclusions about these effects may be less certain than currently accepted.

To illustrate our point, we reanalyze the data from an important recent study on the subject. Despite solid theoretical grounding, early work found mixed evidence that the Democratic Party is able to mobilize more voters in states where there is a strong liberal Democratic Party organization relative to states with a weak organization (Hill and Leighley 1993, 1996; Brown, Jackson, and Wright 1999). Building on Timpone (1998), Brown, Jackson, and Wright (1999) offer the theoretical insight that liberal Democratic Party control of state political institutions should indirectly affect voter turnout through the registration process. Because voting involves a two-stage process in which citizens must register to vote before being allowed to do so, a state Democratic Party organization that wishes to maximize turnout on Election Day must first register potential party supporters to vote.

Brown, Jackson, and Wright (1999) test this hypothesis by regressing state-level registration rates on a measure of liberal party control developed by Hill and Leighley (1996) along with controls for other state-level characteristics that influence voter registration. Their data span four federal elections held from 1984 to 1990 in 45 states. We report the reanalysis of their pooled model in Table 1 (see Brown, Jackson, and Wright 1999, 469, third column of Table 1). The replication of their model, which accounts for heteroskedasticity using the standard White correction, but not clustering, is reported in Column 1. The alternative estimation approaches discussed above—clustering, random effects, HLM, and aggregation to the state level—are reported in the remaining columns.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

Table 1 Reanalysis of Brown, Jackson, and Wright (1999) with and without adjustments for clustering

	(1) Naive OLS	*(2) Clustered OLS*	*(3) Random effects*	*(4) HLM*	*(5) Aggregate OLS*
Liberal party control	1.64 (0.42)	1.64 (0.74)	1.17 (0.68)	1.19 (0.67)	1.70 (0.88)
Registration ease	0.06 (0.01)	0.06 (0.02)	0.04 (0.02)	0.04 (0.02)	0.06 (0.03)
Party competition	−0.03 (0.05)	−0.03 (0.08)	−0.03 (0.06)	−0.03 (0.06)	−0.03 (0.08)
Education	0.40 (0.09)	0.40 (0.15)	0.41 (0.19)	0.41 (0.18)	0.40 (0.20)
State income	−0.02 (0.02)	−0.02 (0.03)	−0.01 (0.02)	−0.01 (0.02)	−0.02 (0.05)
Unemployment rate	0.72 (0.26)	0.72 (0.43)	0.60 (0.21)	0.60 (0.21)	0.79 (0.60)
Mobility	−56.2 (14.0)	−56.2 (24.5)	−58.6 (30.9)	−58.5 (30.3)	−56.8 (31.9)
South	4.02 (1.44)	4.02 (2.52)	3.20 (2.93)	3.23 (2.87)	4.16 (3.06)
Presidential election	3.64 (0.87)	3.64 (0.69)	3.7 (0.43)	3.65 (0.43)	
Constant	45.5 (6.7)	45.5 (11.5)	48.4 (11.0)	48.4 (10.8)	46.3 (15.3)
N	180	180	180	180	45
R-squared	0.39	0.39	NA	NA	0.36
F/χ^2	16.26	37.87	161.2	160.35	2.53

Note. Dependent variable = percent of state voting age population that is registered to vote, 1984–1990 (ranges from 54.5 to 91.3); Liberal party control = Democratic party elite liberalism score × percentage of state legislature controlled by democrats (ranges from 0.19 to 5.09); Registration ease = number of days between the registration closing date and the election × a scale that measures registration difficulty (ranges from 1 to 50); Party competition = level of party competition in state legislature (ranges from 5 to 50); Education = percentage of state population over 25 with a high school education (ranges from 53% to 80%); State income = average state income (ranges from $8777 to $24683); Unemployment rate = state-level unemployment (ranges from 2.4 to 15); Mobility = proportion of state citizens who have lived in current residence for less than two years (ranges from 0.15 to 0.35); South = indicator for form Confederate states; and Presidential election = an indicator for presidential election years. Column 1 replicates the pooled model reported in Column 3 of table 1 in Brown, Jackson, and Wright (1999, 469). See original article for a detailed description of the variables. SEs are given in parentheses. NA, not applicable.

In this application, ignoring the clustered nature of the data leads to SEs that are underestimated by a factor of two for most variables. The SE for the key variable of interest, *liberal party control*, nearly doubles pushing it just beyond the traditional 0.05 threshold for a two-tailed test. Nevertheless, Brown, Jackson, and Wright have good theoretical reasons to expect registration rates to rise when the Democratic Party controls the government during the time period and no reason to expect registration rates to decline, so a one-tailed test is appropriate. *Liberal party control* remains statistically significant across all the models for the one-tailed test of significance. The same is not true for the indicator for southern states. After accounting for clustering, the data no longer provide strong support for the inference that the south is an outlier regarding voter registration rates.

The exception to the inflation of the SEs is the dummy for presidential election years. Clustering SEs *reduces* the SEs by roughly one-quarter compared to those calculated using OLS. The SEs for both random effects and HLM are half as large as those calculated by OLS. Because the unexplained variance in reported rates of voter registration between congressional and presidential elections is greater within a state than the unexplained variance across states during a presidential election, the estimated SEs are reduced. Observing a negative intracluster correlation is rare—especially one of this magnitude. When facing negative intracluster correlations, our impulse is to report the most conservative (i.e., largest) of the estimated SEs so as to avoid Type I errors.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

As expected, there is little difference in the SE estimates among clustered OLS, random effects, HLM, and aggregate OLS. In a few instances, however, aggregate OLS does generate somewhat larger SE estimates. The point estimates for the slope coefficients are also highly similar across the models, but the OLS parameter estimates do differ somewhat from the random effects and HLM regressions. These small differences are likely attributable to unobserved heterogeneity in the residual errors of the random effects and HLM regressions that is correlated with covariates in the model. It is important to remember that all these models make strong assumptions about the absence of unobserved heterogeneity that, if violated, lead to biased point estimates.

3.2 *Political Institutions and Monetary Commitment Regimes*

The need to consider the clustered nature of one's data extends beyond the analysis of mass-level political behavior. Scholars who study economic and government decision making also face instances in which their data are clustered in meaningful ways. Cross-national studies of political institutions offer a prime example because researchers in this area often collect data on multiple countries across time in order to estimate the effects of political institutions. Although this approach is generally sound, it is inappropriate to treat the total number of observations (the number of country-years) as the effective number of observations. Because political institutions and other country-level characteristics change little from year to year, one cannot make the assumption that each country-year observation is independent from all other observations. Furthermore, because country-level variables of interest are often time invariant (or nearly so), it is impossible to include fixed effects *and* estimate the effects of country-level characteristics.

Although we do not wish to address the debate over the necessity of using fixed effects to account for unobserved heterogeneity (Beck and Katz 2001; Green, Kim, and Yoon 2001; Plumper and Troeger 2007), when faced with the task of estimating the effects of time-invariant variables, scholars should adjust their SEs for clustering. To illustrate this point, we reanalyze the data from a recent and influential article on the effects of political institutions on monetary commitment regimes (Broz 2002). In order to keep inflation low, all countries must convince private agents that they will not opportunistically set the value of their currency for short-term gain (e.g., printing more currency than its current value will sustain). In his incisive article, Broz argues that autocracies have a greater need to create highly transparent monetary policy institutions (e.g., fixed exchange rates) to counterbalance the opaqueness of their political institutions. In contrast, democracies—because of their transparent political institutions—can get away with less transparent monetary policy institutions (e.g., independent central banks). Broz tests his hypothesis with data from 123 countries from 1973 to 1995 and uses an ordered probit model to regress monetary policy transparency on a country's level of democracy (as measured in the Polity III data set). He also includes per capita economic growth as a covariate to control for the effects of economic development.

Because Broz is interested in studying the effects of country-level characteristics that change little from year to year, it is impossible for him to include fixed effects in the model. Yet without accounting for the clustered nature of his data, his SEs are smaller than they should be. We replicate Broz's analysis in Table 2. Column 1 presents the original estimates of his baseline model, and Columns 2 and 3 show baseline model estimates that correct for clustering. In Column 2, we employ the standard robust estimation correction for

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

Table 2 Reanalysis of Broz (2002) with and without adjustments for clustering

	(1) No clustering	*(2) Robust clustering*	*(3) Random effects*	*(4) No clustering*	*(5) Robust clustering*	*(6) Random effects*
Lagged DV	1.365	1.365	1.289	1.289	1.289	1.238
	(0.036)	(0.077)	(0.092)	(0.040)	(0.080)	(0.096)
Democracy score	−0.02	−0.02	−0.026	−0.015	−0.015	−0.015
	(0.005)	(0.005)	(0.007)	(0.005)	(0.006)	(0.007)
Per capita GDP	−0.011	−0.011	−0.016	0.023	0.023	−0.03
	(0.005)	(0.007)	(0.011)	(0.009)	(0.010)	(0.013)
Size of economy				−0.239	−0.239	−0.322
				(0.054)	(0.061)	(0.117)
Trade openness				0.169	0.169	0.169
				(0.092)	(0.083)	(0.098)
Inflation differential				−0.306	−0.306	−0.363
				(0.185)	(0.175)	(0.181)
Financial openness				−0.069	−0.069	−0.088
				(0.021)	(0.025)	(0.032)
N	2300	2300	2300	1983	1983	1983
Pseudo-R^2	0.48	0.48		0.47	0.47	
χ^2	2267.12	428.65		2016.08	383.08	

Note. Dependent variable (DV) = monetary policy transparency (1 = free floating, 2 = managed floating, 3 = limited flexibility, 4 = pegged to the dollar); Democracy score = Polity III Democracy score – Polity III Autocracy score (−10 = most autocratic to +10 = most democratic); Size of economy = log of gross domestic product (GDP); Trade openness = (exports + imports)/GDP; Inflation differential = logged absolute difference between country inflation rate and world inflation rate; Financial openness = 14-point scale constructed from International Monetary Fund data by Dennis Quinn (see Broz 2002, 870–73). SEs are given in parentheses.

clustering, and in Column 3, we construct a random effects estimator for an ordered dependent variable using a generalized linear latent and mixed model (GLLAMM; Rabe-Hesketh and Skrondal 2005).

The uncertainty surrounding the point estimate for democracy differs little from the naive model when using clustered SEs (Column 2) but increases by 40% using random effects (Column 3). In neither case is the change sufficient to undermine statistical significance. In contrast, the SEs for per capita gross domestic product (GDP) appear much too small when observations are assumed to be independent. The clustered SE is 40% larger (Column 2), and the random effects SE is more than twice as large (Column 3). What initially appears to be a statistically significant result loses its persuasiveness after accounting for clustering. The two-sided p-value for the coefficient is 0.04 in the naive model, 0.14 in the clustered model, and 0.16 in the random effects model. If effects of economic development were a central concern in this project, clustering would alter the inferences drawn from these data.

In Columns 4 through 6, we reanalyze Broz's with covariate model. The introduction of covariates reduces the degree to which accounting for clustering alters the results. Although the SEs increase for both the democracy and economic development measures, the increases are not large enough to call into question the inferences drawn from the naive model. These findings illustrate a key point mentioned earlier. Collecting informative covariates in the design phase of a study that employs clustered data is essential not just for checking the robustness of the findings—the main rationale offered by most researchers—but to increase the efficiency of the estimates of cluster-level effects.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

4 Conclusion

Our findings underscore the importance of taking into account clustering when estimating group-level effects. Failure to properly do so when analyzing individual data is, in the words of Cornfield (1978, 101), "an exercise in self-deception." Naive SEs overestimate the amount of precision in the parameter estimate and biases *t*-statistics upward. This problem is especially acute when a large portion of the variance in the outcome variable is explained by clustering. Because political science theory often predicts that group-level factors have considerable effects on individual-level decision making, it is imperative that appropriate methods for clustered data are used.

The contribution of our paper is to clear up some confusion over which method is the most efficient. We show that under ideal conditions, clustered SEs, random effects, HLM, and aggregation generate identical estimates (within sampling variability, of course). In doing so, we provide analytical support for similar findings discovered by researchers using simulations (Zorn 2006; Green and Vavreck 2008). We also identify the specific conditions under which researchers gain efficiency from analyzing clustered data at the individual level. Unless the ratio of the within-clusters variance to across-clusters variance of the variable of interest is greater than the variance inflation factor (i.e., the intracluster correlation coefficient weighted by cluster size), researchers do not gain much leverage from analyzing individual-level data. In such instances where this is not the case, does it mean that researchers are better off aggregating individual-level data to the group level? After all, aggregating produces the same precision estimates as the other approaches and it is more transparent.

If one is only interested in estimating the effects of group-level variables *and* is satisfied that conditions for the ecological fallacy are not present, aggregating may prove to be the easiest approach. However, if one is interested in estimating individual-level effects as well or the ecological fallacy is a concern, aggregation makes little sense. In these instances, researchers should choose among the three individual-level approaches reported here. If the number of clusters is plentiful (i.e., above 20), clustered SEs, random effects, and HLM are equally adequate for precision estimates of group-level effects. If there are less than 20 clusters, analysts should avoid using clustered SEs and adopt random effects or HLM. Furthermore, if researchers are also interested in testing whether group-level covariates moderate individual-level effects, HLM may prove to be the most appropriate choice (Steenbergen and Jones 2002).

Moreover, our analysis emphasizes two important research design principles. First, researchers gain efficiency by adding clusters, not by adding individual observations. In settings were the researcher has complete control over data collection, this is an easy design principle to follow. However, few researchers who analyze observational data have the opportunity to increase the number of clusters. In many cases, the number of clusters is fixed by those outside the control of the researcher. There are only 50 states; the Organization for Economic Cooperation and Development only collects data for a subset of countries; and few scholars have direct influence over the sampling design of the American National Election Study. Yet, as Zorn (2006) demonstrates, researchers sometimes have discretion over the *level* at which their data are clustered in the analysis phase. For example, someone studying judicial behavior in state supreme courts can cluster the data by states, judges, or cases. Scholars should make this choice with care and base the decision on the quantity of theoretical interest rather than the method that yields the smallest estimated SEs.

Given limitations in adjusting the number of clusters confronted by many researchers, we cannot stress how much the inclusion of covariates improves the efficiency of point

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

estimates. By collecting individual- and group-level covariates that correlate highly with the outcome variable, researchers can improve the power of their designs without increasing the number of clusters. Even when analysts have little control over the number of clusters, they typically do have some freedom to collect additional covariates. Of course, researchers should include all covariates that are correlated with both the outcome variable and their variable of theoretical interest, but additional efficiency gains can be extracted by including covariates that are simply highly correlated with the outcome variable.

Finally, as always, scholars must appreciate the assumptions underlying the statistical methods they use. When group-level assignment is not random and error components are correlated, these methods will produce biased point and precision estimates. Researchers who analyze observational data should not blindly use these methods and should pay careful attention to the selection process underlying their data.

Funding

Institute for Scholarship and the Liberal Arts at the University of Notre Dame.

Appendix

Proof of Proposition

Comparing equations (4) and (7), individual-level analysis using clustered SEs will offer greater efficiency than aggregated data if and only if the following condition holds:

$$\frac{\sigma^2\left\{1+\left(\bar{n}_g-1\right)\rho\right\}}{\sum_g\sum_i\left(T_{ig}-\bar{T}\right)^2}<\frac{\sigma^2}{\sum_g n_g\left(\bar{T}_g-\bar{T}\right)^2}. \tag{A1}$$

Algebraic manipulations yield the following condition:

$$1+\left(\bar{n}_g-1\right)\rho<\frac{\sum_g\sum_i\left(T_{ig}-\bar{T}\right)^2}{\sum_g n_g\left(\bar{T}_g-\bar{T}\right)^2}. \tag{A2}$$

That is, clustering SEs at the individual level will be more efficient than analyzing the data at the aggregate level when the variance inflation factor, $1+\left(\bar{n}_g-1\right)\rho$, is less than the ratio of the overall variance for the variable of interest to the variance of the variable of interest across clusters. Since

$$\begin{aligned}
\sum_g\sum_i\left(T_{ig}-\bar{T}\right)^2&=\sum_g\sum_i\left(T_{ig}^2-2T_{ig}\bar{T}+\bar{T}^2\right)\\
&=\sum_g\sum_i\left[\left(T_{ig}^2-2T_{ig}\bar{T}_g+\bar{T}_g^2\right)+\left(\bar{T}_g^2-2\bar{T}_g\bar{T}+\bar{T}^2\right)+2\left(T_{ig}\bar{T}_g-T_{ig}\bar{T}-\bar{T}_g^2+\bar{T}_g\bar{T}\right)\right]\\
&=\sum_g\sum_i\left(T_{ig}-\bar{T}_g\right)^2+\sum_g\sum_i\left(\bar{T}_g-\bar{T}\right)^2+2\sum_g\sum_i\left(T_{ig}-\bar{T}_g\right)\left(\bar{T}_g-\bar{T}\right)\\
&=\sum_g\sum_i\left(T_{ig}-\bar{T}_g\right)^2+\sum_g n_g\left(\bar{T}_g-\bar{T}\right)^2.
\end{aligned} \tag{A3}$$

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press

Substituting equation (A3) into (A2) yields:

$$1+\left(\bar{n}_g-1\right)\rho<1+\frac{\sum_g\sum_i\left(T_{ig}-\bar{T}_g\right)^2}{\sum_g n_g\left(\bar{T}_g-\bar{T}\right)^2}. \quad \text{(A4)}$$

That is, individual-level analysis using clustered SEs will increase efficiency over aggregate-level analysis only when the variance inflation factor is less than one plus the ratio of the variance of the variable of interest within clusters to the variance of the treatment across clusters.

If the variable of interest is constant within clusters, $T_{gi}=T_{hj}\,\forall\, i\neq j$ and $g=h$, then $\sum_g\sum_i\left(T_{ig}-\bar{T}_g\right)^2=0$, and there is no possible efficiency gain from individual-level analysis without using covariates.

References

Arceneaux, Kevin. 2005. "Using cluster randomized field experiments to study voting behavior." *Annals of the American Academy of Political and Social Science* 601:169–79.

Beck, Nathaniel, and Jonathan Katz. 2001. "Throwing out the baby with the bath water. A comment on Green, Kim, and Yoon." *International Organization* 55:487–95.

Brown, Robert D., Robert A. Jackson, and Gerald C. Wright. 1999. "Registration, turnout, and state party systems."*Political Research Quarterly* 52(3):463–79.

Broz, J. Lawrence. 2002. "Political system transparency and monetary commitment regimes." *International Organization* 56(4):861–87.

Cornfeld, J. 1978. "Randomization by group: a formal analysis." *American Journal of Epidemiology* 108:100–2.

Donner, Allan. 1998. "Some aspects of the design and analysis of cluster randomized trials." *Applied Statistics* 47:95–113.

Donner, Allan, and Neil Klar. 2000. *Design and analysis of cluster randomization trials in health research*. New York: Arnold Publishers.

Green, Donald P., Soo Yeon Kim, and David H. Yoon. 2001. "Dirty pool." *International Organization* 55:441–68.

Green, Donald P., and Lynn Vavreck. 2008. "Analysis of cluster-randomized field experiments: a comparison of alternative estimation approaches." *Political Analysis* 16(2):138–52.

Greene, William H. 2000. *Econometric analysis, fourth edition*. Upper Saddle River, NJ: Prentice Hall.

Hill, Kim Quaile, and Jan E. Leighley. 1993. "Party ideology, organization, and competitiveness as mobilizing forces in Gubernatorial Elections." *American Journal of Political Science* 37:1158–78.

———. 1996. "Political parties and class mobilization in contemporary United States elections."*American Journal of Political Science* 40:787–804.

Kish, Leslie. 1965. *Survey sampling*. New York: Wiley.

Kmenta, Jan. 1997. *Elements of econometrics: second edition*. Ann Arbor: University of Michigan Press.

Murray, David M. 1998. *The design and analysis of group-randomized trials*. New York: Oxford University Press.

Plumper, Thomas, and Vera E. Troeger. 2007. "Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects." *Political Analysis* 15(2):124–39.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2005. *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical linear models: applications and data analysis methods*. 2nd edn. Thousand Oaks, CA: Sage Publications.

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

Stoker, Laura, and Jake Bowers. 2002. "Designing multi-level studies: sampling voters and electoral contexts." *Electoral Studies* 21:235–67.

Timpone, Richard J. 1998. "Structure, behavior, and voter turnout in the United States." *American Political Science Review* 92:145–58.

Zorn, Christopher. 2006. "Comparing GEE and robust standard errors for conditionally dependent data." *Political Research Quarterly* 59(3):329–41.

https://doi.org/10.1093/pan/mpp004 Published online by Cambridge University Press