

Correctly Modeling Certainty with Clustered Treatments: A Comparison of Methods

Kevin Arceneaux
Assistant Professor
Department of Political Science
Temple University
453 Gladfelter Hall
1115 West Berks Street
Philadelphia, PA 19122
kevin.arceneaux@temple.edu

David W. Nickerson
Assistant Professor
Department of Political Science
University of Notre Dame
217 O'Shaughnessy Hall
Notre Dame, IN 46556
dnickers@nd.edu

March 20, 2007

The authors would like to thank the Institute for Scholarship and the Liberal Arts at the University of Notre Dame for financial support.

Abstract

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. It is well known among statisticians that ignoring the clustered nature of such data underestimates 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 the way in which commonly used methods (clustered robust standard errors, random effects, HLM, and aggregated OLS) estimate the standard errors of group-level effects. We demonstrate analytically and with the help of simulations that under ideal conditions there is no meaningful difference in the standard errors generated by these methods. We also use experimental and observational studies of voter turnout to illustrate the real world performance of these estimators. We conclude with advice for researchers who wish to choose among these methods.

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 statements made by President Carter, analysts risk severely underestimating the uncertainty attached to their causal estimates by ignoring the clustered architecture of their data.

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 of causal estimates. 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 standard errors to model surveying sampling techniques (Kish 1965; Stoker and Bowers 2002), the issue of clustering is given insufficient attention across other areas of research. In this paper,

we compare four practical approaches that researchers can choose to produce a more accurate estimate of their standard errors: clustered robust estimation, random effects, hierarchical modeling, and aggregation to the level of clustering. While 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 tends to focus upon point estimation and largely ignore how uncertainty is modeled.

We attempt to 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 standard errors, we begin with a mathematical explanation of the problem, and provide empirical demonstrations using simulations, a field experiment, and an observational application. Intuitively, one would expect analyzing clustered data at the individual level with appropriate methods (i.e., clustered standard errors, random effects, or HLM) would increase the precision of estimates relative to aggregating to the cluster level. However, we find that under general conditions there is little gain in efficiency by analyzing individual level data, even when the intracluster correlation is very low. Clustering, random effects, hierarchical and aggregation models arrive at very similar standard errors. 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. However, in the conclusion we outline special cases in which scholars should consider adopting one method over another.

Options in Analysis

Basic Problem

Imagine an experiment is conducted with N subjects clustered into G groups, $1 < g < G$,

where each group is comprised of n_g individuals so that $\sum_{g=1}^G \sum_{i=1}^{n_g} i = \sum_{g=1}^G n_g = N$. A treatment T is

applied to a randomly selected set of groups. For the sake of exposition, we examine the case where the treatment is dichotomous (i.e., $T = 1$ for the treatment condition and $T = 0$ for the control group), but the results hold in settings where there are many treatment categories or T is continuous. The empirical goal is to estimate the effect of the treatment, β , on a dependent variable, Y . To begin, consider the basic model

$$Y_{ig} = \alpha + \beta T_g + \varepsilon_{ig} \quad (1)$$

where α is the intercept or average level of the dependent variable for a member of the control group, and ε_{ig} are the idiosyncratic causes of the dependent variable for each individual. Since T is randomly assigned, $\text{cov}(T_g, \varepsilon_{ig}) = 0$. So ordinary least squares (OLS) will yield unbiased point estimates for β , but the question is the degree of uncertainty surrounding our 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 randomization, g .

Conducting the analysis on the individual level assumes N independent observations.

The variance associated with an individual OLS estimate of β is

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

where $\sigma^2 = \sum_g \sum_i \varepsilon_{ig}^2$.¹ However, this formula for the variance assumes the errors to be

independent, which we know to be false from the construction of the experiment. In fact,

$T_{ig} - T_{jh} = 1 \forall g = h, i \neq j$. Thus, a naïve individual level analysis using OLS will overstate the certainty of our estimates.

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

$$\bar{Y}_g = \alpha + \beta \bar{T}_g + \bar{\varepsilon}_g \quad (3)$$

The aggregate analysis is well behaved with $E(\bar{\varepsilon}_g) = 0$ and $\text{var}(\bar{\varepsilon}_g) = \frac{\sigma^2}{n_g}$ (Kmenta 1997, 368).

The variance for OLS estimates of β on the aggregated data is calculated by

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

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

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

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

¹ Typically, $\sigma^2 = \frac{1}{N} \sum_i \sum_g \varepsilon_{ig}^2$. However, the denominator is the variance of T , which is typically estimated

$\text{var}(T) = \frac{1}{N} \sum_i \sum_g (T_{ig} - \bar{T})^2$. For notational ease, the $\frac{1}{N}$ is omitted from both numerator and denominator.

5 nicely illustrates two instructive points. The first lesson is that in most cases where data is clustered into groups, a naïve 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 if the data is aggregated up to the level of randomization.

The second lesson is that if the majority of the variance in the subject population is across groups (i.e., $\sum_g n_g (\bar{T}_g - \bar{T})^2$ 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., $\sum_g \sum_i (T_{ig} - \bar{T}_g)^2$ is small), there is no gain in efficiency from individual-level analysis. The current framework where clusters are assigned treatment conditions and there is no variance within clusters, there is no efficiency gain whatsoever from individual level analysis because $\sum_g \sum_i (T_{ig} - \bar{T}_g)^2 = 0$. Adding covariates to the analysis or varying the treatment 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., $\sum_g \sum_i (T_{ig} - \bar{T}_g)^2$ is large), there would appear to be efficiency gains from moving from the aggregate to the individual level. 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.

Clustered Standard Errors

The first method makes no changes to the estimation procedure from OLS, but adjusts the standard errors to account for correlation between subjects 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, $E(\varepsilon_{ig}^2) = \sigma^2 = \sigma_\gamma^2 + \sigma_u^2$. By assumption, the group component is assumed to be uncorrelated across groups (i.e., $E(\gamma_g \gamma_h) = 0 \forall g \neq h$). The individual level error component is assumed to be independent both within and across cluster. That is, $E(\varepsilon_{ig} \varepsilon_{jh}) = 0 \forall g, h \text{ \& } i \neq j$.

Thus, the model presented in equation 1 can be re-written as

$$Y_{ig} = \alpha + \beta T_g + \gamma_g + u_{ig} \quad (6).$$

The critical concept in the analysis is the ratio of variance within clusters to overall variance in the model. This concept is typically referred to as the intraclass 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 intraclass correlation implies that the clusters explain little and that the effective sample size is closer to the number of subjects, 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 intraclass correlation means the individual-level OLS estimates will be severely biased and in need of adjustment. The formula for adjusting the variance estimates for β to account for clustering is presented in equation 7 (see Donner 1998, 98 or Murray 1998, 362).³

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

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.

$$\text{Var}(\beta_{\text{Clustered}}) = \frac{\sigma^2 \{1 + (\bar{n}_g - 1)\rho\}}{\bar{n}_g G \sum_g \sum_i (T_{ig} - \bar{T})^2} \quad (7)$$

Assuming that intracluster correlation is always non-negative, a condition that applies to nearly every applied setting, $1 + (\bar{n}_g - 1)\rho \geq 0$.⁴ Thus, the first lesson to take from equation 7 is that naïve 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. Equation 7 also illustrates why adding subjects 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. Furthermore, Donner and Klar (2000) demonstrate that clustered standard errors provide overly optimistic standard errors when the size of clusters is small. In most 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 subject 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 $1 + (\bar{n}_g - 1)\rho$ 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 condition:

⁴ The term $1 + (\bar{n}_g - 1)\rho$ was termed the design effect by Kish (1965), but variance inflation factor is more popular in the medical sciences (e.g., Donner 1998).

$$\text{Var}(\beta_{\text{Clustered}}) < \text{Var}(\beta_{\text{Aggregated}}) \Leftrightarrow 1 + (\bar{n}_g - 1)\rho < \frac{\sum_g \sum_i (T_{ig} - \bar{T})^2}{\sum_g (\bar{T}_g - \bar{T})^2} \quad (8)$$

That is, efficiency is gained from examining the individual-level data if and only if the variance inflation factor, $1 + (\bar{n}_g - 1)\rho$, is smaller than the ratio of the overall variance in the treatment to the variance of mean treatment across groups. Since both methods account for the clustered nature of the data, large differences in precision are not anticipated in most circumstances.

Random Effects

The model presented in equation 6 is structurally identical to the well-developed random effects model used in panel data settings. Typically random 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 formulae 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. Simulation results confirm this intuition.

Hierarchical Modeling

On the surface, the multi-level 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. Hierarchical linear models (HLM) capture the full complexity of multi-level data by explicitly modeling both the individual level and the clustered level (see

Steenbergen and Jones 2002 for a helpful introduction). Equation 9 presents the group randomized trials in an HLM model:

$$\begin{aligned} Y_{ig} &= \alpha + u_{ig} \\ \alpha &= \mu + \beta T_g + \gamma_g \end{aligned} \tag{9}.$$

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

$$Y_{ig} = \mu + \beta T_g + \gamma_g + u_{ig} \tag{10}.$$

Equation 10 is obviously 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, while naïve individual-level OLS results will consistently underestimate uncertainty in clustered data, clustered standard errors, random effects, and hierarchical models all adequately account for the structure of the data. The simulations in the next section confirm this contention.

Simulations

To demonstrate the properties of the different approaches to estimating standard errors discussed in the last section, we simulate clustered data so that the data generation process is known. To highlight the downward bias of naïve OLS standard errors in the presence of clustering, we examine conditions with low ($icc = 0.01$), medium ($icc = 0.1$), and high ($icc = 0.3$) intracluster correlations. An intracluster correlation of 0.3 may appear artificially

high, but such levels routinely occur in studies of families (e.g., Stoker and Jennings 2005, p. 56) and cross-country studies (e.g., Steenbergen and Jones 2002, p. 231). Regardless, it will be demonstrated that downward bias in the standard errors is a serious problem even in the presence of very low *ICC*. We also manipulate the number of clusters, dividing 1,000 observations into 10 and 50 groups, in order to test the small cluster properties of the alternate estimation techniques.

In the simulations, clusters were randomly assigned to a dichotomous treatment group with $p = 0.5$ to mimic an experimental setting in which the error structure of the data should be well behaved. We generated an outcome y in the following way:

$$y = \beta T + \lambda_1 X_g + \lambda_2 X_{ig} + \gamma_g + u_{ig}, \quad (11)$$

where, X_{ig} and X_g are individual and cluster-level covariates respectively. Both X_{ig} and X_g conform to a standard normal distribution, but our results are robust to distribution assumptions. Similarly, the cluster-level, γ_g , and the individual-level, u_{ig} , errors are drawn from normal distributions where the standard deviations are manipulated to generate the desired *ICC*. The coefficients were set to $\beta = 2$, $\lambda_1 = 0.3$, and $\lambda_2 = 0.3$, but none of our conclusions are sensitive to the choice of coefficients. Simulations were run 1,000 times using the computer code in the Appendix.

The standard errors of the estimated treatment effect, β , for each technique in the simulation results are presented in Table 1. The estimated coefficients are not presented because the point estimates should not and do not vary from technique to technique. That is, clustering, in and of itself, does not bias point estimates. Problems only arise when the cluster-level errors are correlated with the treatment of interest (i.e., $\text{cov}(\gamma_g, T_g) \neq 0$). The standard deviation of the simulated distribution of the estimated treatment effect coefficient is presented in the column

labeled “Empirical SE” and represents the true level of variation and uncertainty in the parameter estimates. The estimated standard error for each technique (presented in columns 5-9) should be compared to this benchmark.

These simulations make plain six key points. First, ignoring the clustered nature of the data severely underestimates the standard errors of the treatment effect. Naïve regression very consistently posits a level of certainty regarding the results that is unwarranted. Second, the bias grows as the intraclass correlation increases. The empirical standard errors are twice as big as naïve OLS standard errors at low levels of *ICC* and three or even seven times bigger at higher levels of *ICC*.⁵

In contrast, the other techniques correct for this downward bias equally well (lesson #3). As expected, clustered robust standard errors, random effects, HLM, and simple aggregation all produce roughly the same standard errors when there are a large number of clusters. However, the fourth lesson is that, as mentioned above, robust clustered standard errors are biased downwards when the number of groups is small. The typical rule of thumb cut-off provided by the medical literature is that 20 clusters are sufficient for reliable estimates and, nothing in our simulations leads us to disagree with this point.

Adding clusters not only makes robust clustered standard errors accurate, but also increases power (lesson #5). While the total number of observations in each simulation is held constant (1,000), organizing those observations into 50 clusters instead of 10 doubles statistical precision at most levels of intraclass correlation. Unfortunately, researchers are unable to manipulate the number of clusters in many experimental settings and almost never in most observational studies. Thus, we also suggest that researchers plan to collect informative

⁵ A corollary of lesson #2 is that reliable inference is more difficult as intraclass correlation grows. The simulations help to demonstrate that the standard errors for estimates increase steadily as the correlation between observations within clusters increases.

covariates to increase statistical power (lesson #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 standard errors for the treatment of interest using naïve 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.

One exception to this rule should be noted. In the vast majority of cases, robust clustered standard errors, random effects, HLM, and simple aggregation will yield more conservative standard errors than naïve OLS. While theoretical instances of negative intracluster correlation are extraordinarily rare, the *ICC* might be weakly negative for a particular sample. In these instances, the standard errors for naïve OLS may be larger than the techniques that take into account clustering. As a simple rule, researchers should report the larger of the naïve OLS standard errors and the cluster sensitive techniques.

In the next two sections, we demonstrate that these principles extend beyond the artificial world of computer simulations with both experimental and observational data. The experimental example conforms to the assumptions underlying the math we presented above. Specifically, each cluster has an equal chance of receiving the treatment, allowing the error terms to be uncorrelated with treatment assignment. Consequently, it offers a glimpse of how the various estimators showcased here perform under ideal conditions. However, many political scientists do not work under ideal conditions, so we use observational data to extend our analysis to an instance where critical assumptions may be violated.

An Experimental Example: Voter Mobilization by Precinct

In this example, we use data collected from a randomized field experiment designed to test the effect that door-to-door get out the vote (GOTV) canvassing has on voter turnout. The experiment took place in 2003 in Kansas City, MO where a community organization sought to boost support for a municipal bond proposal to increase funding for mass transportation. The group targeted 28 precincts where they believed heavy consumers of mass transportation lived. For practical reasons, the researcher assigned precincts, rather than the 9,712 registered voters who resided in these precincts, to treatment ($n = 14$) and control groups ($n = 14$). After the election, official voting records were obtained for all individuals within the 28 precincts and used to measure turnout (for more details on the study see Arceneaux 2005).

Because registered voters are nested within precincts and randomization occurred at the precinct level, these data illustrate the perils of ignoring clustering when estimating the standard error for the treatment effect in individual-level analyses. We estimate the treatment effect with same four approaches: naïve OLS, OLS with clustered standard errors, random effects, HLM, and OLS aggregated to the precinct level. We also estimate the treatment effect with and without covariates included in the model. The Kansas City voter file has a rich amount of information about subjects' past voting behavior with voting history records for 21 previous elections spanning a seven year period (1996-2003). By itself, vote history explains approximately 32 percent of the variance in 2003 turnout (adjusted- $R^2 = 0.319$). Because previous voting behavior is a strong predictor of turnout, including vote history in these models should markedly increase the efficiency of the treatment effect estimate.

The results, which are shown in Table 2, strongly echo our simulations. Naïve OLS biases the standard errors downward by two to three times. Even though the intraclass correlation is miniscule for these data ($ICC = 0.017$), failure to take into account the clustered

nature of the data produces misleading estimates of uncertainty for causal effects. Furthermore, the standard error estimates are roughly the same across all four approaches that take into account clustering. It appears that the clustered standard errors are slightly larger than those for the other approaches when no covariates are included, while the standard errors for the aggregated OLS are slightly larger than the others when covariates are included. However, these are small differences, and bolster our contention that all three approaches generate roughly equivalent standard errors, which may differ slightly across different samples. In no instance do these slight differences affect statistical inference. We can safely reject the null hypothesis that there is no treatment effect with a one-tailed test under all four approaches.

However, we can markedly improve the efficiency of our treatment effect estimate by including relevant covariates. As the results in row 2 clearly demonstrate, the standard errors in the analyses with covariates included are approximately 66% smaller than those in the analyses without covariates. The additional precision from covariates cannot be understated. Without covariates, a researcher would need to triple the number of clusters to achieve comparable efficiency gains. In this particular instance, the neighborhood association conducting the canvassing was unwilling to partition precincts or increase the number of precincts included in the experiment, so the number of clusters is fixed in this study. Given the number of instances in both experimental and observational settings where the number of clusters is beyond the control of the researcher, collecting relevant covariates should not be an afterthought.

Finally, note that the point estimates do not differ significantly across the various models and estimators. This is possible because treatment assignment is orthogonal to the covariates and the individual- and group-level error terms are uncorrelated. If this assumption is violated, researchers may very well get different point estimates across these methods. As we discuss in

the next example, researchers should be concerned with more than just the estimates of standard errors when analyzing clustered observational data.

An Observational Example: State Registration Laws and Turnout

Over 100 articles have been written on the effects of voter registration laws on turnout in the past 25 years. These studies often use state-level survey data, such as the Current Population Study (CPS), to measure turnout at the individual level and estimate individual-level regression analyses with measures of state registration laws included to estimate their effects on individuals' propensity to vote. There are a number of reasons to suspect that the voting behavior of individuals is more correlated within states than across. Not only do voters in a state share a similar culture, they face the same political conditions and personalities that are difficult to model. For these reasons, analysts should take into account state-level clustering when estimating the effects of voter registration laws on turnout.

To illustrate this expectation, we revisit Wolfinger and Rosenstone's (1980) seminal study by reanalyzing the 1972 CPS. Our goal here is not to critique this important study or even challenge its results. Given limitations in computing power and the undeveloped state of methodological approaches for clustered data 25 years ago, it would hardly be fair to expect Wolfinger and Rosenstone to address these issues in their original study. Instead, our aim is purely pedagogical, and to this end, our analysis of the 1972 CPS differs somewhat from their study. First, we include all 84,000 observations for which there are valid responses on all the covariates. Second, we augment the registration law model reported in the earlier work (see Wolfinger and Rosenstone 1980, Appendix F, 129) with a broader set of demographic variables. As in the original, we estimate the effects of registration laws by including measures for the closing date (i.e., the number of days before the election that a resident must register to vote), an

indicator for whether the registration office kept irregular hours, an indicator for whether the registration office was open on evenings, weekends, or both, an indicator for whether the state did not allow absentee registration, and the number of hours the polls were open on Election Day. In addition to these covariates, we include a dummy variable for the presence of a gubernatorial election, the south, and a spate of demographic variables (age, age-squared, married, white, black, female, education, education-squared, and income). Third, in order to be consistent with the preceding discussion, we use a linear probability model.

The results are reported in Table 3. The naïve OLS estimates reported in column 1 show strong effects for state registration laws. An early closing date reduces turnout and extended registration office hours boosts turnout by 4 percentage points. Oddly, disallowing absentee registration boosts turnout and longer poll hours reduces turnout. However, these findings change when we take into account the clustered nature of the data. The standard errors for the state-level variables are larger than the naïve OLS estimates across all of the clustered models, thereby changing our interpretation of the coefficients. By more fully accounting for model uncertainty, the strange effects for absentee registration and poll hours become statistically insignificant, and the effect for extended office hours becomes less robust. In this application, taking into account clustering affects statistical inference.

These results reinforce what the simulations and experimental example illustrate. Nevertheless, there are some notable differences. While the estimates for some of the state-level slope coefficients and standard errors are quite similar across clustered OLS, random effects, HLM, and aggregated OLS, this is not true for all of the parameter estimates. The effects of irregular office hours and absentee registration switch signs in the random effects and HLM analyses suggesting these estimates are not robust. In the random effects and HLM models, the

standard errors for extended registration hours are more than three times larger than the clustered OLS estimate and almost double the size of the aggregated OLS estimate. Because state registration laws are not randomly assigned, as our group-level variables have been up to this point, it is possible that key assumptions underlying these methods are violated, leading to bias in the point estimates and standard errors.

Moreover, the coefficients for the demographic covariates are highly significant across the individual-level analyses, but are almost all statistically insignificant in the aggregate-level analysis. These null results are likely accounted for by the low level of variation in some of these demographic variables across the states and the small degrees of freedom available, illustrating an important point. To the extent researchers are also interested in the effects of individual-level variables on dependent variables measured at the individual level, aggregation makes little sense.

Conclusion

Our simulations and analyses 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.” Naïve standard errors 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 dependent 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. The surprising answer is that under ideal conditions, clustered standard errors,

random effects, HLM, and aggregation generate identical estimates (within sampling variability, of course). It does not appear that researchers gain much leverage from analyzing individual-level data, even when the intracluster correlation coefficient is quite small. Does this 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 straightforward.

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 standard errors, 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 standard errors 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, we cannot stress how much the inclusion of covariates improves the efficiency of point estimates. This point is especially relevant for researchers who are designing cluster-randomized experiments. By collecting individual- and group-level covariates that correlate highly with the dependent variable, researchers can improve the power of their designs without increasing the number of clusters.

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

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.
- Cornfeld, J. 1978. "Randomization by group: A formal analysis." *American Journal of Epidemiology* 108:100-102.
- 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.
- 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.
- 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* 21:235-267.
- Stoker, Laura and M. Kent Jennings. 2002. "Political Similarity and Influence between Husbands and Wives." In *The Social Logic of Politics*, ed. Alan S. Zuckerman. Philadelphia: Temple University Press.
- Wolfinger, Raymond and Steven Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.

Table 1: Simulation Results

Groups	ICC	Covariates	Empirical SE	Naïve OLS	Cluster	RE	HLM	Aggregate
10	0.01	No	0.24	0.07	0.21	0.23	0.23	0.23
		Yes	0.10	0.07	0.08	0.09	0.10	0.10
	0.1	No	0.28	0.06	0.24	0.27	0.27	0.27
		Yes	0.19	0.06	0.15	0.18	0.18	0.20
	0.3	No	0.36	0.05	0.30	0.35	0.35	0.35
		Yes	0.30	0.05	0.24	0.29	0.29	0.32
50	0.01	No	0.11	0.06	0.11	0.11	0.11	0.11
		Yes	0.06	0.06	0.06	0.07	0.07	0.07
	0.1	No	0.13	0.06	0.13	0.13	0.13	0.13
		Yes	0.09	0.05	0.09	0.09	0.09	0.09
	0.3	No	0.16	0.05	0.16	0.16	0.16	0.16
		Yes	0.13	0.05	0.12	0.13	0.13	0.13

Each simulation was run 1,000 times.
Numbers reported are standard errors.

Table 2: The Effect of Door-to-Door Canvassing on Voter Turnout, 2003 Kansas City Field Experiment

	Naïve OLS	Clustered OLS	Random Effects	HLM	Aggregate OLS
Without Covariates	4.4** (0.9)	4.4* (2.5)	4.2* (2.4)	4.1* (2.4)	4.4* (2.4)
With Covariates	5.4** (0.8)	5.4** (1.6)	5.2** (1.1)	5.0** (1.6)	6.0** (1.8)
Number of Observations			9712		
Number of Clusters			28		
ICC			0.017		

Note: Point estimates are intent-to-treat effects and the standard errors in parentheses.
 ** $p < 0.01$, * $p < 0.05$, one-tailed test.

Table 3: The Effects of State Registration Laws on Reported Turnout, 1972 CPS

Variable	Naïve OLS	Clustered OLS	Random Effects	HLM	Aggregate OLS
Age	0.0168** (0.0005)	0.0168** (0.0013)	0.0168** (0.0005)	0.0168** (0.0005)	-0.0863 (0.0629)
Age ²	-0.0001** (0.0000)	-0.0001** (0.0000)	-0.0001** (0.0000)	-0.0001** (0.0000)	0.0010 (0.0006)
Married	0.0244** (0.0038)	0.0244** (0.0059)	0.0244** (0.0038)	0.0244** (0.0038)	-0.0233 (0.3272)
White	0.2386** (0.0139)	0.2386** (0.0499)	0.2598** (0.0150)	0.2614** (0.0151)	0.2111 (0.1307)
Black	0.2675** (0.0148)	0.2675** (0.0524)	0.2931** (0.0158)	0.2950** (0.0159)	0.1475 (0.1650)
Female	-0.0059+ (0.0031)	-0.0059 (0.0056)	-0.0052+ (0.0031)	-0.0051+ (0.0031)	-2.1384** (0.5648)
Education	0.0355** (0.0021)	0.0355** (0.0036)	0.0356** (0.0021)	0.0356** (0.0021)	-0.2082 (0.1289)
Education ²	0.0002* (0.0001)	0.0002 (0.0001)	0.0002* (0.0001)	0.0002** (0.0001)	0.01+ (0.0054)
Income	0.0228** (0.0007)	0.0228** (0.0014)	0.0232** (0.0007)	0.0232** (0.0007)	0.0306 (0.0200)
South	-0.0679** (0.0035)	-0.0679** (0.0129)	-0.0585** (0.0154)	-0.0558** (0.0180)	-0.0733** (0.0234)
Closing Date	-0.0017** (0.0002)	-0.0017* (0.0007)	-0.0022** (0.0008)	-0.0023** (0.0009)	-0.0014+ (0.0008)
Irregular Registration Office Hours	0.0051 (0.0040)	0.0051 (0.0137)	-0.0226 (0.0165)	-0.0234 (0.0194)	0.0112 (0.0138)
Evening/Weekend Office Hours	0.0437** (0.0077)	0.0437** (0.0143)	0.0170 (0.0522)	0.0172 (0.0624)	0.0583+ (0.0297)
No Absentee Registration	0.0074* (0.0033)	0.0074 (0.0107)	-0.0017 (0.0151)	-0.0020 (0.0179)	-0.0003 (0.0130)
Hours Polls Open	-0.0003** (0.0001)	-0.0003+ (0.0002)	-0.0001 (0.0004)	-0.0001 (0.0004)	-0.0002 (0.0004)
Gubernatorial Election	0.0131** (0.0040)	0.0131 (0.0135)	0.0307+ (0.0166)	0.0317 (0.0196)	0.0085 (0.0150)
Constant	-0.6707** (0.0222)	-0.6707** (0.0612)	-0.6753** (0.0335)	-0.6773 (0.0372)	3.8541* (1.6104)
Observations	84,384	84,384	84,384	84,384	51
R ²	0.14	0.14	0.14	NA	0.83
ICC	0.012				

Note: Standard errors in parentheses.

** $p < 0.01$, * $p < 0.05$, + $p < 0.10$, two-tailed test

Appendix: STATA 9.2 Code for Simulations

```
*Simulations
set mem 100m
set more off

program define better, rclass
    version 8
    syntax [, obs(integer 1) g(real 0) icc(real 0)]
    drop _all
    set obs `obs'
    gen u_i = invnorm(uniform())
    gen u_g = invnorm(uniform())
    gen byte group = group(`g')
    bysort group: egen counter = seq()
    replace u_g = . if counter~=1
    bysort group: egen t_g = median(u_g)
    replace u_g = t_g
    drop t_g

    gen treatment = uniform()
    replace treatment = 1 if treatment>0.5
    replace treatment = 0 if treatment<0.5
    replace treatment = . if counter~=1
    bysort group: egen trt = median(treatment)
    replace treatment = trt
    drop trt

    gen xi = invnorm(uniform())
    gen xg = invnorm(uniform())
    replace xg = . if counter~=1
    bysort group: egen tx = median(xg)
    replace xg = tx
    drop tx

    gen y = 2*treatment + 0.3*xi + 0.3*xg + `icc'*u_g + (1-`icc')*u_i

    regress y treatment
    scalar beta_ols_nc= _b[treatment]
    scalar se_ols_nc= _se[treatment]

    regress y treatment xg xi
    scalar beta_ols_wc= _b[treatment]
    scalar se_ols_wc= _se[treatment]

    regress y treatment, cluster(group)
    scalar beta_rcse_nc= _b[treatment]
    scalar se_rcse_nc= _se[treatment]

    regress y treatment xg xi, cluster(group)
    scalar beta_rcse_wc= _b[treatment]
    scalar se_rcse_wc= _se[treatment]

    xtreg y treatment , re i(group)
    scalar beta_re_nc= _b[treatment]
    scalar se_re_nc= _se[treatment]
    scalar rho_nc = e(rho)

    xtreg y treatment xi xg, re i(group)
    scalar beta_re_wc= _b[treatment]
    scalar se_re_wc= _se[treatment]
    scalar rho_wc = e(rho)

    xtmixed y treatment || group:, iterate(100)
```

```
scalar beta_hlm_nc= _b[treatment]
scalar se_hlm_nc= _se[treatment]

xtmixed y treatment xi xg || group:, iterate(100)
scalar beta_hlm_wc= _b[treatment]
scalar se_hlm_wc= _se[treatment]

collapse y treatment xg xi, by(group)
reg y treatment
scalar beta_ag_nc= _b[treatment]
scalar se_ag_nc= _se[treatment]

reg y treatment xg xi
scalar beta_ag_wc= _b[treatment]
scalar se_ag_wc= _se[treatment]

end
```