STATISTICAL INFERENCE FOR THE LINEAR MODEL WITH CLUSTERED DATA

Jeffrey J. Harden

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Political Science (American Politics).

Chapel Hill
2009

Approved by:

Thomas M. Carsey

Skyler J. Cranmer

James A. Stimson

ABSTRACT

JEFFREY J. HARDEN: Statistical Inference for the Linear Model with Clustered Data
(Under the direction of Thomas M. Carsey.)

Political scientists often confront clustered data, which can present problems for statistical inference. Through Monte Carlo simulation I examine the performance of standard error methods in clustered data for two linear estimators: Ordinary Least Squares (OLS) and Median Regression (MR). I consider changes to several parameters: sample size, number of clusters, intra-cluster correlation, and error term distribution (normal, which favors OLS as the most efficient estimator, and Student's t, which favors MR). Results indicate that conventional OLS and MR standard errors are often, but not always, biased downward in clustered data. Within OLS, the performance of the robust cluster standard errors (RCSE), which are designed for clustered data, is conditional on the level of covariate variation and the severity of cluster correlation. Regarding MR, two nonparametric methods perform well. I conclude that researchers should carefully examine the nature of the clustering in their data before choosing a standard error method.

ACKNOWLEDGMENTS

I greatly appreciate the contributions made by several people over the course of this project. Don Green and Lynn Vavreck provided me with complete access to their Monte Carlo code. Robbie Hogan was very generous in sharing his data. Bruce Desmarais fielded several questions and helped with R code. Shubin Liu provided great assistance in efficiently using the research computing resources at UNC. Finally, I greatly appreciate the input from committee members Skyler Cranmer and Jim Stimson and the many pages of comments and countless conversations with my advisor, Tom Carsey.

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vi

Introduction 1

- The Problem of Clustering 2
- OLS and MR 4
- Standard Error Calculation Methods 5

Monte Carlo Simulations 7

- Objectives 7
- Experimental Model Estimation 9
- The Simulation Procedure 10

Results 11

- Number of Clusters 11
- Sample Size 13
- Intra-cluster Correlation 14
- Error Term Distribution 16

Discussion 18

An Application: Clustered Data in State Politics Research 19

- Original Results 20
- Accounting for Clustering 21
- MR Replication 23

Conclusions 24

APPENDIX 26

- Computational Information 26
- Coverage Probability 26
- Sample R code 27

REFERENCES 30

LIST OF TABLES

Table

1 Monte Carlo Simulation Dynamics 9

2 Independent Variables in Hogan (2008, Table 2) 20

3 OLS and MR Models Predicting Challenger Spending as a Percentage of Incumbent Spending (Hogan 2008, Table 2) 21

LIST OF FIGURES

Figure

1 Normal and Student's $t(3)$ Distributions 8

2 Effects of Increasing the Number of Clusters on OLS and MR Standard Error Methods for $N = 1{,}200$, Normal ε, and $\rho \equiv 0.1$ 12

3 Effects of Increasing the Sample Size on OLS and MR Standard Error Methods for β_2 (cluster-level variable), Normal ε, and $\rho \equiv 0.1$ 14

4 Effects of Changing $\rho \equiv 0.001$ on OLS and MR Standard Error Methods with Normal ε 16

5 Effects of Changing the Distribution of ε to a $t(3)$ on OLS and MR Standard Error Methods for β_2 (cluster-level variable) with $\rho \equiv 0.1$ 17

6 OLS and MR Models of Hogan (2008, Table 2) with Largest Confidence Interval Estimates 24

Introduction

Political science research must often account for problems that arise in the data generating process. One such difficulty that is common in several subfields is the possibility of unmodeled correlation between observations within groups, commonly referred to as "clustering" or "mixed-level data." Comparative scholars, for example, face this issue when studying the effects of country-level variables on individuals. Similarly, within American politics observations might be clustered in Congressional districts, states, or other smaller units. Several studies show that clustering creates a downward bias in the standard errors of Ordinary Least Squares (OLS) regression coefficients, leading to a higher likelihood of committing a Type I error—rejecting the null hypothesis when it is in fact true (e.g., Cornfield 1978; Moulton 1990; Wooldridge 2002, 2003; Arceneaux 2005).

One approach to solving this problem is to adjust the standard errors to account for clustering. In political science, estimating "robust cluster standard errors" (RCSE) for the coefficient estimates has become popular in recent years, likely due to the ease in which the method can be implemented in several statistical software programs. However, despite the mathematical logic that justifies the use of RCSE, it is necessary to assess their effectiveness, along with that of other standard error methods, in an experimental setting. Recent work suggests that the RCSE method may not be as useful as conventional wisdom indicates (Green and Vavreck 2008; Arceneaux and Nickerson n.d.).

In the present study, I provide such an assessment within the context of linear regression. First, I offer a systematic test of the effects of clustering on several standard error estimation techniques for two different linear estimators: OLS and Median Regression (MR). In so doing, I vary conditions such as the number of clusters in the data, the sample size, and the level of correlation within clusters. Furthermore, to compare OLS and MR objectively, I vary whether the error term is normally distributed or not. While OLS assumes normality, MR is often a more efficient linear estimator when the normality assumption is violated (Bassett and Koenker 1978; Koenker and Bassett 1978; Desmarais and Harden n.d.). I show that accounting for the effects of clustering on a linear regression model involves not only choosing the correct method for estimating standard errors, but also choosing the correct regression estimator.

The Problem of Clustering

Several disciplines besides political science, including labor economics, education, and various medical fields commonly encounter clustered data (e.g., Donner and Wells 1986; Moulton 1990; Ukoumunne 2002; Wooldridge 2002, 2003; Donner and Klar 2004). Political science is especially subject to the issue in part because a broad question studied in the discipline is how higher-level features of political systems influence the behavior of actors at lower levels. For instance, comparative and international relations research often examines observations grouped in sub-regional, sub-national, or dyadic units (e.g., Posner 2004; Golder 2006; Kasara 2007; Crescenzi 2007; Danilovic and Clare 2007; McDonald 2007; Büthe and Milner 2008). Similarly, within the American context researchers commonly encounter observations grouped in smaller units such as states, media markets, counties, cities, or households (Tolbert, McNeal and Smith 2003; Barreto, Segura and Woods 2004; Carson and Crespin 2004; Wolfinger, Highton and Mullin 2005; Arceneaux and Huber 2007; Arceneaux and Nickerson 2009). Importantly, additional scholarship indicates that not accounting for the clustered nature of the data can pose problems for substantive conclusions, even when there is no omitted variable bias (Carsey and Wright 1998; Zorn 2006; Primo, Jacobsmeier and Milyo 2007; Green and Vavreck 2008).

The Design Effect

In an ideal setting, observations for an analysis would be selected and assigned to treatment and control randomly. This would allow the researcher to assume independence among all observations in the study. However, political science almost never enjoys this level of control. In statistical terms, clustering in the data constitutes a violation of the assumption of independent and identically distributed errors (i.i.d.). This violation creates a "design effect" which generates downward bias in the standard errors (Kish 1965; Donner and Klar 2000). Formally, the measure of the degree to which individuals within clusters are similar to each other is calculated as the value ρ. This value is defined by

$$\rho = \frac{s^2_{between}}{s^2_{between} + s^2_{within}} \tag{1}$$

where $s^2_{between}$ is the average variance of the error term across clusters and s^2_{within} is the average variance within clusters (Green and Vavreck 2008). The design effect combines ρ with the average number of individuals within each cluster to yield a measure of the bias to conventional statistical inference methods for a given sample of data. Formally, the design effect is calculated as

$$DE = 1 + \rho(m - 1) \tag{2}$$

where m is equal to the average number of observations within each cluster (Donner and Klar 2004). From this, conventional OLS standard errors will be biased downward by a factor of $\sqrt{1 + \rho(m - 1)}$. For instance, as Green and Vavreck (2008, 143) report, if the design effect is 4, conventional OLS standard errors will be biased downward by a factor of 2 and thus the threshold for statistical significance at the 0.05 level will be $\pm$ 3.92 for a two-tailed test hypothesis test rather than $\pm$ 1.96.

Approaches to Dealing with Clustered Data

To this point, most research on clustered data has assumed that the robust cluster technique adequately estimates the magnitude of the design effect by reporting results in which the RCSE are larger than the OLS SE (e.g., Carsey and Wright 1998; Arceneaux 2005; Zorn 2006). A key focus of this study is whether even the larger RCSE are providing an accurate estimate of coefficient variability. If they are not, then it may be more beneficial for researchers to use another estimation strategy to deal with clustered data.[1]

However, despite the fact that alternatives do exist, dealing with clustering by adjusting the standard errors is still quite common. Within political science, several studies use some form of robust standard error whenever the presence of clustering seems to exist (e.g., Carsey and Jackson 2001; Branton 2004; Buckley and Westerland 2004; Bonneau 2005; Stratmann 2006; Gabel and Scheve 2007; Berry 2008; Gerber, Green and Larimer 2008).[2] Based on this increasing popularity, the goal of the present study is to provide a more rigorous evaluation of the standard error choices available.[3] Put differently, I assess the extent to which adjusting standard errors is actually accommodating clustered data.

[1] Several methods are available for analyzing clustered data other than the robust standard error adjustment procedure examined here. Another involves aggregating the data up to the level of the clustering. This removes the correlation among within-cluster residuals, but it does so at the cost of degrees of freedom and theoretical richness. As a result, it is usually more advantageous for a researcher to keep the unit of analysis at the lowest level, and adjust the model to accommodate the clustered data. This can be done in several ways, including Generalized Least Squares (GLS) (Greene 2002), Generalized Estimating Equation (GEE) models (Zeger and Liang 1986; Zorn 2006), or Hierarchical Linear Models (HLM) (Raudenbush and Bryk 2002; Gelman and Hill 2007). However, it should also be noted that all of these methods have their own limitations (see Steenbergen and Jones 2002; Zorn 2006; Green and Vavreck 2008; Arceneaux and Nickerson n.d.).

[2] Zorn (2006, 332) reports that the number of articles in the three major political science journals containing the term "robust standard error" increased from a minimum of one in 1992 to a maximum of 11 (*Journal of Politics*) in 2002. An informal survey of these journals in the period since 2002 indicates a continuation of this trend.

[3] Not all of the studies cited here use OLS in their analyses. For simplicity, I only consider linear regression in my evaluations.

OLS and MR

In addition to testing OLS standard errors, I also consider standard error methods for MR, which is methodologically similar to OLS, but more efficient under certain conditions. A typical OLS model fits the data according to the conditional-mean of the dependent variable (Y) at each fixed value of an independent variable (X), or $E[Y|X]$. Similar logic applies in a MR model, except that the response is fit according to a conditional-median value.

In any linear regression, model residuals are produced as the predicted values subtracted from the actual values, or $Y_i - \hat{Y}_i$. OLS minimizes the sum of the squared residuals:

$$\min \sum (Y_i - \hat{Y}_i)^2. \tag{3}$$

In contrast, MR minimizes the sum of the *absolute deviations*—instead of squaring the residuals, the method simply sums their absolute values. Through one of two methods[4] , this approach fits the regression line such that the number of points above the line is equal to the number below the line:

$$\min \sum |Y_i - \hat{Y}_i|. \tag{4}$$

Both OLS and MR yield linear regression coefficients as a solution to a minimization problem.[5] The least squares method assumes the error term, ε, is normally distributed with mean zero and constant variance. MR assumes a Laplace distribution for the errors, which displays heavier tails than the normal. Desmarais and Harden (n.d.) show that a linear model will be estimated more efficiently by either OLS or MR depending on how close the empirical distribution of the errors is to each of these assumed distributions. OLS is a more efficient method when the error term is closer to a normal and MR is more efficient when it is closer to a Laplace. Examples of heavy-tailed distributions that favor MR include the Cauchy and low-degree of freedom Student's t. In the present study, I examine the consequences of manipulating the error term to favor the efficiency of one method over the other for the performance of the standard errors.

[4] A common approach to solving for MR parameter estimates is to use the Simplex Algorithm (see Koenker 2005). Alternatively, a Bayesian approach to the problem utilizes Markov chain Monte Carlo (MCMC) with a Laplace-distributed likelihood function (Yu and Moyeed 2001).

[5] Put differently, just as OLS can be characterized as maximum likelihood estimation (MLE) with a normally-distributed likelihood function, MR can also be characterized as MLE with a Laplace-distributed likelihood function.

Standard Error Calculation Methods

OLS

I conduct this analysis by considering three methods for calculating OLS standard errors: the conventional standard errors (OLS SE), RCSE, and bootstrapping (BSE). In a typical linear model such as $Y = X\beta + \varepsilon$, the OLS variance-covariance matrix is calculated as

$$\text{var-cov}(\hat{\beta}) = (X^T X)^{-1} \cdot X^T \mathbf{\Phi} X \cdot (X^T X)^{-1} \tag{5}$$

where $\mathbf{\Phi} = \varepsilon\varepsilon^T$. If ε is homoskedastic and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j$, then $\mathbf{\Phi} = \sigma^2\mathbf{I}$ and Equation 5 reduces to $\sigma^2(X^T X)^{-1}$. OLS SE can then be calculated as the square-roots of the elements on the main diagonal of the variance-covariance matrix. Of course, this assumes that the errors are i.i.d., an assumption violated when clustering is present.

To account for this, RCSE constitute a simple adjustment to Equation 5 in which it is assumed that $\mathbf{\Phi} \neq \sigma^2\mathbf{I}$. In this case, the elements on the main diagonal of $\mathbf{\Phi}$ are not constrained to the same value (i.e., not constrained to σ^2) and off-diagonal elements *from observations within the same cluster* are not constrained to be zero, as they are in the conventional calculation. Rather, covariances between observations within the same cluster are estimated empirically from the residuals (see Williams 2000; Green and Vavreck 2008, 142).

This is essentially a slight modification to the procedure of the well-known Huber-White robust standard errors, which are designed to correct for heteroskedasticity (Huber 1967; White 1980, see also Long and Ervin 2000). As with RCSE, the Huber-White method estimates the variance-covariance matrix such that $\mathbf{\Phi} \neq \sigma^2\mathbf{I}$. However, while the elements on the main diagonal of $\mathbf{\Phi}$ are not constrained to the same value, the elements of the off-diagonals (covariances) are *all* constrained to zero in the Huber-White method, as in the standard calculation of the variance-covariance matrix. By incorporating the possibility of non-zero covariances within clusters, RCSE are designed to be robust to heteroskedasticity *and* cluster correlation. However, the assumption that observations from different clusters are independent remains (Williams 2000).

Another approach to calculating standard errors is through a re-sampling technique, such as bootstrapping (Efron 1979, 1982). This method estimates the parameters repeatedly by drawing from the sample data with replacement. After a sufficiently large number of replications, the distribution of parameter estimates can be used to generate a variance-covariance matrix. This approach uses the empirical distribution of the bootstrap replications to approximate the asymptotic variance of the parameters. The procedure has been shown to perform well under violations to the i.i.d. assumption (Liu 1988).

MR

Like in OLS, several methods exist for calculating MR standard errors.[6] Here I consider three: the conventional asymptotic standard errors (ASE), a kernel density estimate (KSE), and bootstrapping (BSE). The ASE method is roughly analogous to the method for calculating the conventional OLS SE (Hao and Naiman 2007). The MR asymptotic variance-covariance matrix takes the form

$$\frac{p(1-p)}{N} \cdot \frac{1}{f_{\varepsilon^p}(0)^2} \cdot (X^T X)^{-1} \tag{6}$$

where $p = 0.5$ in the case of MR, N is the sample size, and $f_{\varepsilon^p}(0)^2$ is the density of the error term.[7] Thus, as in OLS, the matrix is calculated as the product of a scalar and $(X^T X)^{-1}$, but in the MR case the multiplier, $\frac{p(1-p)}{N} \cdot \frac{1}{f_{\varepsilon^p}(0)^2}$ is an estimate of the asymptotic variance of the error term evaluated at the sample median (Hao and Naiman 2007, 45). Once the variance-covariance matrix is calculated, standard errors are then constructed in the usual way, as the square-roots of the elements on the main diagonal. The crucial issue with the ASE method is the same as that of the conventional OLS variance-covariance matrix: it assumes the errors are i.i.d. (Koenker 2005; Hao and Naiman 2007). As is the expectation with the OLS SE, this reliance on asymptotics should cause the ASE to be biased downward under clustering in the data.

The other two MR methods I consider are nonparametric and neither are explicitly designed to accommodate cluster- and individual-level variance like the RCSE. However, they also do not explicitly assume i.i.d. errors, which should allow them to be more robust to clustering. The BSE method for MR is identical to the method for OLS described above. Bootstrapping does not impose a distribution on the data and has been shown to accommodate non-i.i.d. data, and thus it may provide better coverage probability than the ASE.

The KSE method, first introduced by Powell (1991), uses a kernel approximation to construct standard errors. It estimates the sampling distribution of the MR model by creating a kernel smooth over the model residuals, then calculates the variance-covariance matrix as the variance of the multivariate kernel density function. This method should be more robust to clustering because it estimates standard errors based on the sample data rather than asymptotic assumptions.

---

[6] Derivations of several methods can be found in Koenker and Bassett (1978), Powell (1991), Koenker (1994), and Koenker and Machado (1999). Analyses of these methods include Rogers (1992), Gould (1992), and Koenker and Hallock (2001).

[7] Equivalently, the ASE can be described as the inverse hessian of the MR likelihood function.

Monte Carlo Simulations

Next I describe a simulation procedure designed to generate a data set with a random cluster component in the error term, estimate the model repeatedly, and assess the differences in the standard errors calculated by each method. The code used to generate the data structure is based on that of Green and Vavreck (2008); an example is provided in the Appendix.

Objectives

In these simulations I assess the extent to which a 95% confidence interval created from each standard error method actually includes the true parameter estimate in 95% of the repeated Monte Carlo simulations. A standard error that is closer to this 95% standard is considered "better" than one that is further away. This process is described in more detail below.

I conduct the experiment under changing data conditions. First, I simulate data with 5, 10, 25, 40, and 50 clusters. The literature on clustering indicates that adding clusters improves standard error accuracy (e.g., Killip, Mahfoud and Pearce 2004; Arceneaux 2005; Green and Vavreck 2008; Arceneaux and Nickerson n.d.). To test this, I hold the number of observations constant, and divide the data into increasingly more clusters—each with its own unique random effect—to assess how standard error estimates behave. This adds to what is often called the "effective sample size." The maximum cluster value of 50 is chosen as a realistic reflection of the number of clusters (i.e., states or countries in a region) most scholars in political science can expect.

Next, I consider changes to the sample size by conducting the experiment with 200, 800, and 1,200 observations. These values are selected to reflect realistic sample sizes of studies in political science. Note that sample size is closely related to ρ. Consider the formula for the design effect (Equation 2); holding ρ constant, the magnitude of the downward bias to OLS SE will increases as the sample size increases, because more observations will be added to the same number of clusters (see Killip, Mahfoud and Pearce 2004).[8]

Another important parameter to examine dynamically is the value of ρ itself. Donner and Klar (2004) explain that the effects of adjustments to other parameters—such as sample size or number of clusters—are dependent on ρ. If the level of intra-cluster correlation is low, the

[8] This may seem counterintuitive given the common expectation in social science that increasing sample size improves precision. Consider the $\varepsilon\varepsilon^T$ matrix from a linear model estimated on clustered data. This matrix is filled with a diagonal of ones, (assumed) zeros in off-diagonal elements corresponding to observations from different clusters, and ρ in off-diagonal elements corresponding to observations within the same cluster. For a constant number of clusters, as observations are added more cells take on the value ρ and fewer take on the value zero. Thus, the matrix begins to look less like the assumed $\mathbf{\Phi} = \sigma^2\mathbf{I}$. This should correspond to deteriorating performance by the conventional standard errors.

improvement to standard error accuracy from adding clusters is relatively modest. However, if ρ is large, adding more clusters has a bigger effect (Donner and Klar 2004, 420). For this reason, I conduct the experiments at two values of ρ: 0.001, which is common in medical experiments (Donner and Klar 2004), and 0.1, which is the value used in the Green and Vavreck (2008) study.

Finally, I consider the implications of violations to the normality assumption by conducting the experiment with a normal error and one drawn from a Student's t distribution with three degrees of freedom. The t distribution with infinite degrees of freedom is equivalent to the normal, but when it has low degrees of freedom, it takes on heavier tails such that a sample will likely contain more outliers than a sample from a normal. Figure 1 provides a visual comparison of the two distributions. The figure plots the theoretical densities of a normal distribution with $\mu = 0$ and $\sigma = 1$ and a t distribution with three degrees of freedom.

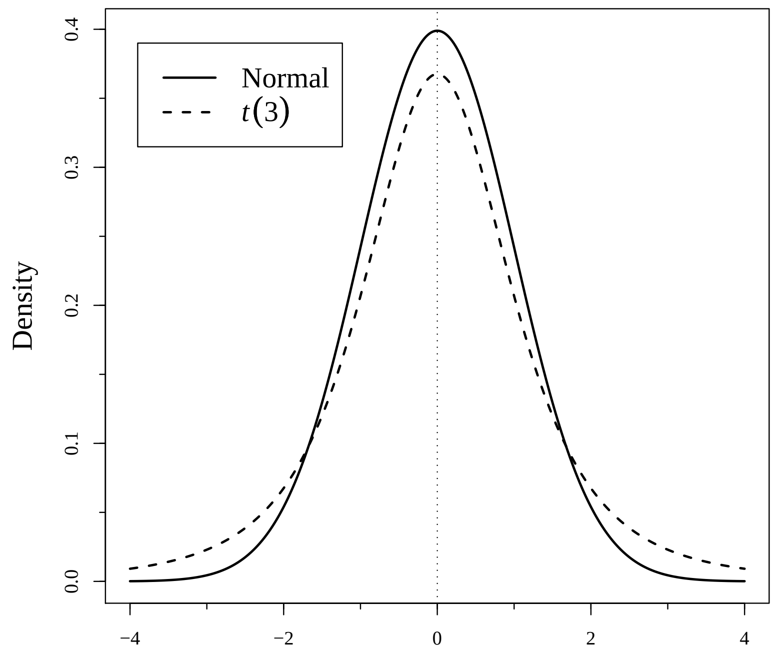

Fig. 1: Normal and Student's $t(3)$ Distributions

Desmarais and Harden (n.d.) show that MR is a more efficient linear estimator than OLS for a model with an error term drawn from a low degree of freedom t distribution. This is due to the fact that the OLS estimator responds exponentially to deviations from the conditional-mean, while MR responds linearly to deviations from the median. In other words, MR will be more resistant to the random outliers produced in a draw from the $t(3)$. Thus, half of the experiment is set up to favor OLS and half is set up to favor MR. When the error is normal, the OLS standard error methods should perform better and when the error is drawn from a $t(3)$, the MR methods should perform better.[9] Table 1 summarizes the dynamic aspects of the study.

Table 1: Monte Carlo Simulation Dynamics

Variable Name	Description
Number of Clusters	5, 10, 25, 40, and 50 clusters
Sample Size	200, 800, and 1,200 observations
ρ	Assumed values: 0.001, 0.1
Distribution of ε	Normal and $t(3)$

Experimental Model Estimation

To assess the effect of these parameters I construct the following experimental model with dependent variable Y, independent variables X_1 and X_2, and an error term ε. These variables are indexed by N individual observations $i \in (1, \ldots, N)$ and C clusters of observations $c \in (1, \ldots, C)$. X_1 varies at the individual level and X_2 is a cluster-level variable. I assess standard error performance on both variables.

$$Y_{ic} = \alpha + \beta_1 X_{1ic} + \beta_2 X_{2c} + \varepsilon_{ic} \tag{7}$$

Following Green and Vavreck (2008), the model is defined such that $\alpha = 0$, $\beta_1 = 0.85$, and $\beta_2 = 0.5$, although these values can be changed without affecting results. In addition, the error term (ε) is broken into two components: an individual-level disturbance e_i and a cluster-level disturbance υ_c such that

$$\varepsilon_{ic} = e_i + \upsilon_c. \tag{8}$$

As mentioned above, in half the experiment e_i and υ_c are distributed normally and in half they are distributed according to a t with three degrees of freedom. Additionally, the error term is uncorrelated with the independent variables to avoid creating bias in the coefficient estimates (Gujarati 2003, 71).

The next step is to select a method for evaluating and comparing the standard error methods. This process is not as straightforward as evaluating different methods for calculating coefficient estimates because the smallest standard errors are not necessarily the correct standard errors. Several methods exist in the literature, including calculating "overconfidence" percentages (Beck

[9] This design slightly favors OLS because it is the MLE for a normal distribution. As mentioned above, the equivalent distribution for MR is the Laplace. However, current packages in R do not allow for random sampling from a bivariate Laplace distribution, which is necessary for the Monte Carlo simulation (see lines 74–82 in the example code provided in the Appendix). The $t(3)$ does favor MR, but not to the extent of the Laplace (see Desmarais and Harden n.d.).

and Katz 1995), or comparing the standard deviation of the vector of simulated coefficient estimates with the average standard error (Green and Vavreck 2008).

Another method is to calculate a coverage probability (Newcombe 1998; Bradlow, Wainer and Wang 1999; Platt, Hanley and Yang 2000; Ukoumunne 2002). This involves constructing 95% confidence intervals from the standard errors produced by each simulation and calculating the proportion of these confidence intervals that includes the true parameter. The expectation is that this proportion should be 0.95 if the standard error method is "correct."[10] A value less than 0.95 indicates a downward bias (i.e., toward Type I errors) and a value greater than 0.95 indicates a conservative (Type II error) bias. I use this method because it is common in the statistics literature and because of its simple interpretation, but results are not dependent on this choice. For example, evaluating standard error performance by comparing the mean standard error from the Monte Carlo replications to the standard deviation of the coefficient estimates from the same replications produces the same substantive conclusions. See the Appendix for a more detailed description of the method.

The Simulation Procedure

The Monte Carlo simulation procedure unfolds as follows with N observations and C clusters. At the three different values of N (200, 800, and 1,200) the procedure is repeated 10,000 times at each value of C (5, 10, 25, 40, 50). This yields a total of 300,000 simulations.[11] A single replication of the simulation procedure operates in the following way:

1. A clustered dataset is constructed according to Equations 7 and 8 with N observations and C clusters.[12]
2. OLS and MR models are fit to the data.
3. From the OLS model, the coefficient estimates, OLS SE, RCSE, and BSE are extracted. From the MR model, the coefficient estimates, ASE, KSE, and BSE are extracted.
4. The process is repeated until 10,000 replications have been performed.

[10] Results remain the same if another standard, such as 0.50, is set as the confidence interval level.

[11] See the Appendix for computational information on the simulation procedure.

[12] I use the same technique as Green and Vavreck (2008) to impose a specified value of ρ on the data. This involves setting the variance of the *population* from which the cluster-level random effect is drawn, which subjects the empirical value, $\hat{\rho}$, to sampling error. At $\rho \equiv 0.001$, the average value of $\hat{\rho}$ was 0.003 with a standard deviation of 0.005 and at $\rho \equiv 0.1$, the average value of $\hat{\rho}$ was 0.097 with a standard deviation of 0.025. Although these values are slightly off from their intended targets, the ordering is consistent. In other words, the smaller population value of ρ produces empirical values that are smaller than the larger population value. For more on calculating ρ, see Donner and Wells (1986), Ukoumunne (2002), or Donner and Klar (2004).

At each value of N and C, this process yields 10,000 coefficient estimates for each parameter in the OLS and MR models and 10,000 estimates for each standard error method and each coefficient. After storing this information, the final step is to calculate the coverage probability of each standard error method.

Results

The results of the simulations indicate that each of the various dynamics outlined in Table 1 affect standard error performance. I present the results graphically across the range of clusters in the simulations. In each graph, the number of clusters is plotted on the x-axis and the coverage probability of each standard error method is plotted on the y-axis. A dashed line is drawn at 0.95 to indicate the standard of what a "correct" 95% confidence interval should cover.

Number of Clusters

Previous studies have found that standard errors are generally biased downward when fewer clusters are present, and improve as the number of clusters increases. Figure 2 provides some support for this finding. The graphs plot the coverage probabilities for the OLS and MR standard error estimates for β_1 (left panels) and β_2 (right panels) with $N = 1{,}200$ observations, a normal error, and $\rho \equiv 0.1$.

Most notably, these results indicate a difference between how the various methods perform on the individual-level variable and the cluster-level variable and a difference between the MR and OLS standard error methods. First, consider the OLS results in the top-left and top-right panels of Figure 2. For β_1, the coefficient on the individual-level variable (top-left panel), the OLS SE and BSE are nearly perfect—both fall on or just around the 0.95 line across the range of clusters. The lines are so close in this and many subsequent graphs that it is difficult to differentiate the BSE line from the OLS SE line. The RCSE, however, are consistently biased downward. This bias is fairly substantial—about 10 percentage points—at 5 clusters, though it decreases to less than 1 percentage point as the number of clusters increases.

In contrast to OLS β_1, the results for β_2, the cluster-level OLS variable (Figure 2, top-right panel), indicate that the RCSE perform the best. All three of the methods improve as the number of clusters increases in that graph, with the RCSE taking on the highest values across the range. Interestingly, however, all three of the methods exhibit a downward bias under these conditions, with the RCSE coverage probability ranging from 70% to approximately 93% and the other two methods reaching only about 89% at their highest values.

Next, in the MR results, the standard errors estimated for β_1 (Figure 2, bottom-left panel) are all quite close to the 0.95 level and perform equally well across the range of clusters. The ASE and BSE are essentially directly on the line, which makes differentiating the two difficult, while the KSE exhibit a conservative bias; they are too large by about 2 percentage points. Moving to the bottom-right panel of Figure 2, the results for β_2 indicate that the MR standard errors behave similarly to those of the OLS model. Again, the upward curve exists as the number of clusters increases. In this case, the KSE perform the best, though all three exhibit a downward bias, especially at low cluster values. Indeed, with 5 clusters, the KSE method coverage probability is only about 65% and the BSE and ASE methods produce confidence intervals that cover at about 60%.

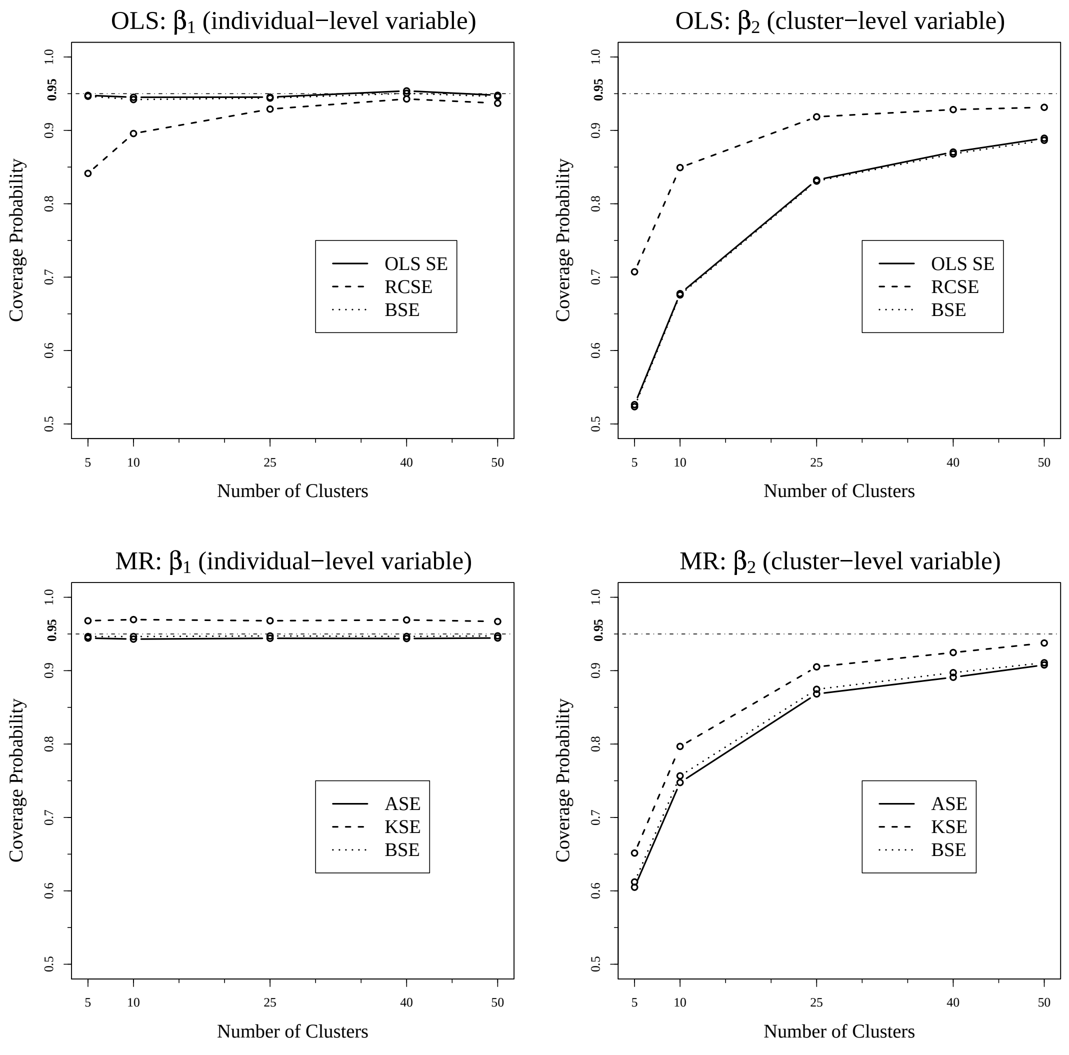

Fig. 2: Effects of Increasing the Number of Clusters on OLS and MR Standard Error Methods for $N =$ 1,200, Normal ε, and $\rho \equiv 0.1$

Sample Size

Common intuition of standard errors would suggest that adding observations would improve performance. However, within the context of clustered data, Figure 3 shows that an increase in sample size increases the magnitude of the design effect. The graphs depict the coverage probabilities for the OLS and MR standard error estimates for β_2 under the same conditions as the right panels of Figure 2, but with the smaller sample sizes of $N = 200$ (left panels) and $N = 800$ (right panels). Results for β_1 are not shown because there is virtually no difference in performance at different sample sizes for that coefficient.

At the smallest sample size of 200 (Figure 3, top-left panel), all three OLS methods are quite similar; they exhibit the familiar trend upward along the x-axis and all are biased downward (though only by about 1–2% after 25 clusters). In addition, the results indicate that the conventional OLS SE are the best method under these conditions. However, moving to the top-right panel of Figure 3 shows that this result changes at a sample size of 800. Again, all three methods show a downward bias, but the RCSE are notably better than the other two methods, even by about 3 percentage points at 50 clusters. Thus, a sample size increase magnifies the downward bias of the OLS SE and BSE but does not alter the performance of the RCSE as much.[13]

The MR results are slightly different. The simulations performed with a sample size of 200 (Figure 3, bottom-left panel) show the familiar trend upward across the range of clusters. However, a larger separation is evident. No method hits the 0.95 level perfectly. While the BSE are the closest, they are biased slightly downward by about 9 and 4 percentage points, respectively, at 5 and 10 clusters. Additionally, the KSE fall farther away from the 0.95 line, but are generally too large by about 2–3%.

Moving to the sample size of 800 (Figure 3, bottom-right panel), the picture changes slightly. The relative ordering remains the same, with the KSE falling at the highest coverage probabilities across the range of clusters. However, in this case even the KSE exhibit a downward bias at low cluster values, including an extreme of about 73% coverage at 5 clusters. The ASE and BSE are consistently biased downward, though they are only 2–3 percentage points off from the 95% level at 50 clusters.

[13] Comparing the top-right panel of Figure 3 to the top-right panel of Figure 2 indicates that RCSE performance is similar at 800 and 1,200 observations.

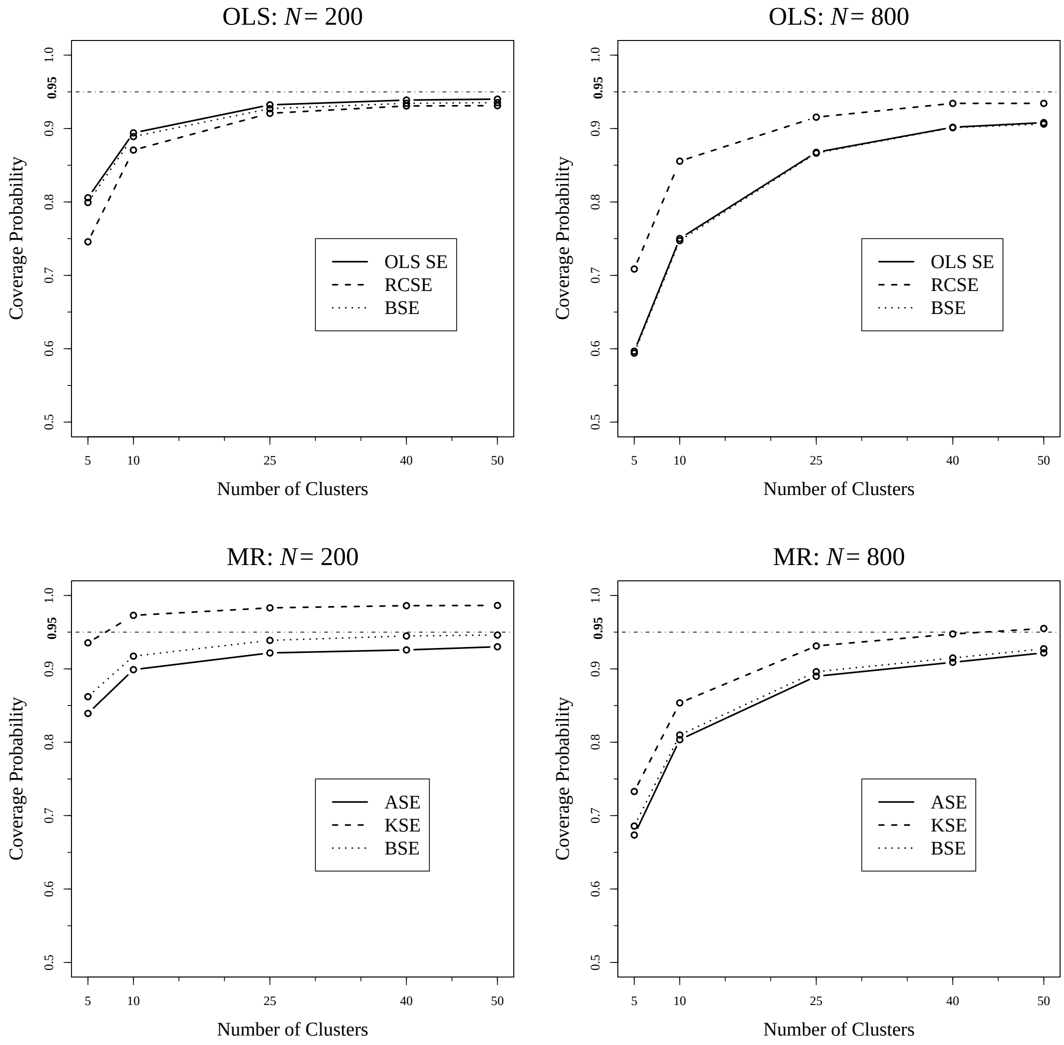

Fig. 3: Effects of Increasing the Sample Size on OLS and MR Standard Error Methods for β_2 (cluster-level variable), Normal ε, and $\rho \equiv 0.1$

Intra-cluster Correlation

Next I consider how changing the level of correlation within clusters influences standard error performance. The left panels of Figure 4 show the results under identical conditions to those in the left panels of Figure 2 (β_1, $N = 1{,}200$, normal error), but with $\rho \equiv 0.001$ instead of 0.1. This means that although observations are grouped in clusters, the violation of the i.i.d. assumption is less severe. Similarly, the right panels of Figure 4 shows the simulation results under identical conditions to those in the right panels of Figure 3 (β_2, $N = 800$, normal error), but with $\rho \equiv 0.001$ instead of 0.1.

The top-left panel of Figure 4 indicates that changing ρ does not have a substantial impact on the OLS standard error methods for the individual-level variable. The RCSE are still biased downward by between 10 and 1 percentage points as the number of clusters increases, and the near-perfect performances of the OLS SE and BSE remain. However, the OLS β_2 results tell a much different story. Recall from the top-right panel of Figure 3 that at a sample size of 800 and $\rho \equiv 0.1$, the RCSE perform notably better than the other OLS methods for β_2. In contrast, when $\rho \equiv 0.001$ (top-right panel of Figure 4), the RCSE are slightly biased downward at levels of about 19 to 2 percentage points across the range of clusters, while the OLS SE and BSE perform almost precisely at the 0.95 level. In other words, changing the value of ρ changes which standard error method is most accurate when estimated for a cluster-level OLS variable. For a small value of ρ, the conventional OLS SE or bootstrapping are more accurate than the RCSE, but as ρ increases, the RCSE perform better.

The MR results are similar. Figure 4 shows virtually no change to the MR standard error performance for the individual-level variable: the bottom-left panels of Figures 2 and 4 are nearly identical. All three lines are very near the 0.95 level, although the BSE fall the closest to the standard. Next, the bottom-right panel of Figure 4 shows that reducing the value of ρ improves all three MR methods when estimated for a cluster-level variable. The downward bias under small numbers of clusters is eliminated and all three methods perform at or near the 0.95 level. The BSE method again falls closest to the line with the KSE method covering at about 1–2% higher than the 95% level.

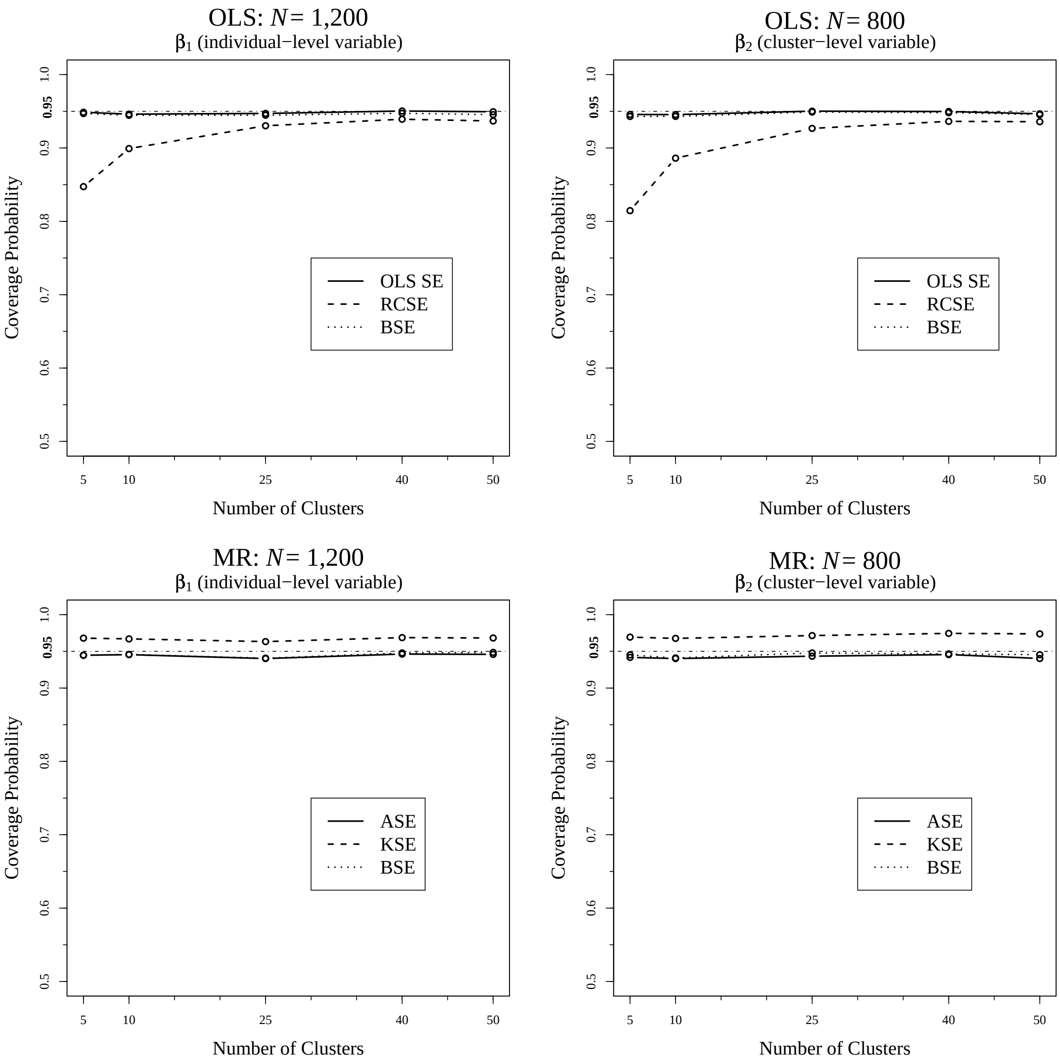

Fig. 4: Effects of Changing $\rho \equiv 0.001$ on OLS and MR Standard Error Methods with Normal ε

Error Term Distribution

Finally, I consider the implications of changes to the error term distribution on these standard error methods. Recall that only results with a normal error have been presented to this point, and the MR standard error methods have performed comparably well to those of OLS. The results in Figure 5, however, show that violations to the normality assumption can be problematic for OLS. The top and bottom panels, respectively, plot OLS and MR β_2 results for a sample size of 200 (left panels) and 800 (right panels) with a $t(3)$ error distribution.

The top panels show that OLS suffers considerably at both sample sizes. For instance, compare the top-right panel of Figure 5 to the top-right panel of Figure 3. The only difference between those two graphs is the distribution of ε. Although the BSE and RCSE both perform better under the $t(3)$ specification than the OLS SE, all three are sharply biased downward when the error term is not normal. The RCSE, which perform the best of the three methods, do not even reach the 90% level across the range of clusters.

Two of the three MR methods are less affected by the $t(3)$ error distribution. The bottom panels of Figure 5 show that while the ASE are consistently too small for both sample sizes, the other two methods perform reasonably well, though the increased design effect is quite evident when moving from 200 observations to 800. The KSE appear to be the best choice. They are within 1 percentage point of the 95% standard for most of the range of clusters when the sample size is 200, and are slightly less biased than the RCSE with a sample size of 800, reaching a coverage probability of about 92% with 50 clusters. In comparison, the OLS RCSE only reach 89% coverage under identical conditions.

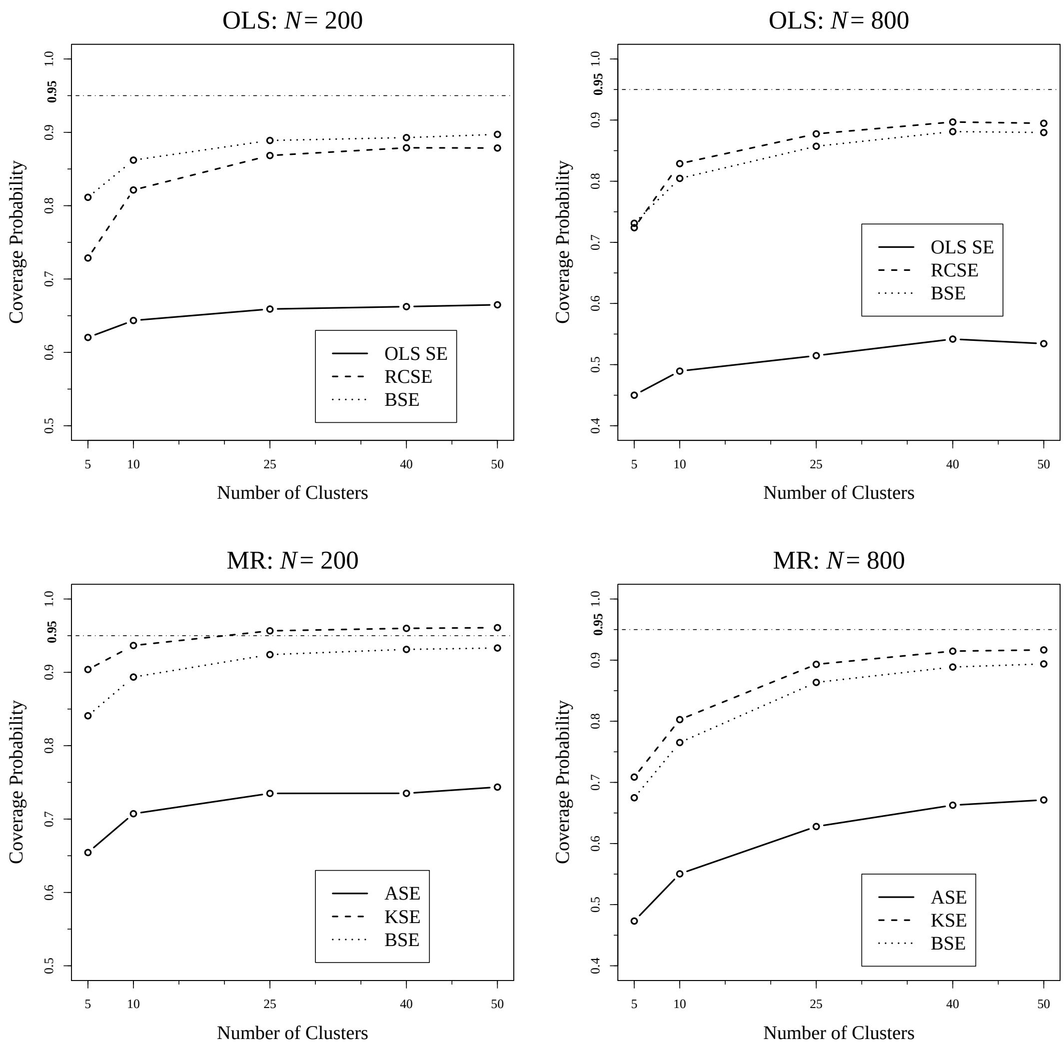

Fig. 5: Effects of Changing the Distribution of ε to a $t(3)$ on OLS and MR Standard Error Methods for β_2 (cluster-level variable) with $\rho \equiv 0.1$

Discussion

The simulation results show that each of the parameters examined dynamically can influence the performance of these standard error methods. Before addressing these factors, however, it is important to note briefly that the results clearly indicate that adding more clusters to the data is beneficial. Virtually all methods improve or remain constant as the effective sample size is increased. This finding is consistent with previous work on clustering (e.g., Arellano 1987; Moulton 1990; Wooldridge 2002, 2003; Killip, Mahfoud and Pearce 2004; Arceneaux 2005; Green and Vavreck 2008; Arceneaux and Nickerson n.d.).

The results also show a consistent difference in standard error performance depending on the level of the variable for which the standard errors are estimated. For lower-level variables—those that vary at the observational level—the OLS SE outperform the RCSE. In fact, under a normal error, the only instances in which the RCSE actually do improve standard error estimates are for the cluster-level variable. The MR results do not show a clear-cut difference in this regard. The KSE are often the best choice for both levels of variable, though they do tend to be slightly conservative when estimated for lower-level variables.

As previous findings would suggest, an increase in the sample size contributes to increasing bias in the the conventional standard errors for both estimators. Adding more observations magnifies the design effect created by the correlation within clusters. In contrast, the RCSE method performs better than the OLS SE, but still generates biased standard errors when the number of clusters is small. When more clusters are present in the data, the additional residuals provide more information with which to estimate cluster correlation. However, while the RCSE method may be improving inference for the cluster-level variables in a model, it may at the same time be estimating standard errors for individual-level variables that are too small.

Within the MR context, the BSE perform the best at the smallest sample size, but the design effect biases them downward at the larger sample sizes. Though the KSE are slightly too conservative at 200 observations, they display more resistance to the design effect, and actually perform as well or better than the RCSE estimated for the OLS model even when the error term is drawn to favor OLS. For instance, consider the right panels of Figure 3 ($N = 800$, β_2, normal error, $\rho \equiv 0.1$). While both the KSE and RCSE are biased downward at the lower cluster values, the KSE fall closer to the 0.95 standard at 25, 40, and 50 clusters (KSE: 0.931, 0.948, 0.955; RCSE: 0.916, 0.934, 0.934).

Next, this study shows that the level of correlation within clusters substantially affects standard error performance. For instance, the RCSE show a clear improvement from the other two OLS methods with $\rho \equiv 0.1$ (see Figure 3), but not with $\rho \equiv 0.001$ (see Figure 4). The MR results suggest some changes in this regard as well. While the KSE perform at or near the 0.95 level at $\rho \equiv 0.1$, they become slightly too conservative while the BSE perform quite well at $\rho \equiv 0.001$. As expected due to their theoretical reliance on the i.i.d. assumption, the ASE are consistently the most biased MR option in the study. Interested readers who download the replication materials will find that further increasing ρ leads to all of the methods examined here producing estimates that are increasingly biased downward (although the RCSE display the most resistance to increased levels of ρ). Thus, accounting for group-level variation via standard error adjustment may only be feasible at smaller values of ρ.[14]

Finally, the results indicate that the distribution of the error term is crucial to selecting the appropriate standard error method. Comparing Figures 2–4, with Figure 5 shows a stark contrast between the OLS methods with and without a normal error. While one or more methods perform at close to the 0.95 level under normality, all three OLS methods exhibit a downward bias under the $t(3)$ error term.

In contrast, the MR standard error methods reflect the robustness of the estimation technique, as they are less affected by changes to the distribution of ε. While the ASE are biased downward under both specifications (and more so under the $t(3)$), the KSE and BSE still perform fairly well. This finding suggests that if a researcher decides to use the standard error adjustment method in dealing with clustered data, one of the first choices made should be that of selecting the correct linear estimator. Desmarais and Harden (n.d.) provide a sample-based test for determining whether OLS or MR is the more efficient linear estimation technique.[15]

An Application: Clustered Data in State Politics Research

Next I apply the information learned from the Monte Carlo study to existing research in the state politics literature. In "Policy Responsiveness and Incumbent Reelection in State Legislatures," Hogan (2008) examines how the voting behavior of state legislators can influence their chances of reelection. He looks at this process in three specific areas: the decision of challengers to run against incumbents, fundraising success by both challengers and incumbents, and votes received by challengers and incumbents. The data used in the analysis come from approximately

[14] Additional simulation work suggests that HLM provides better standard error estimates than pooled regression with RCSE when the value of ρ is greater than 0.2.

[15] This procedure will eventually be available through the R package `elm`.

2,686 incumbents in both the lower and upper houses of 14 states in 1996 and 1998.[16]

In the current analysis, I focus on replicating the OLS model predicting challenger spending as a percentage of incumbent spending (Hogan 2008, Table 2, 866). The main independent variable, *Partisan Policy Position*, is a measure of incumbent divergence from expected district preferences on economic and regulatory policy. This is constructed by regressing legislative voting scores from the National Federation of Independent Business (NFIB) on several district-level measures of demographics, such as income, racial makeup, and education levels. The absolute value of the residuals from this equation are used as a measure of how severely a given incumbent's voting record diverges from the preferences of the district. A larger residual signifies a legislator whose voting record is strongly divergent from district preferences. Within the context of the challenger spending model, the expectation is that greater divergence will correspond to increased spending by the challengers—an incumbent out of touch with his or her district will draw more significant opposition. The model also includes several control variables, as described in Table 2.

Table 2: Independent Variables in Hogan (2008, Table 2)

Variable Name	Description
Partisan Policy Position	High values indicate divergence toward party base, away from district median
Political Party	1 = Democrat, 0 = Republican
Major Party Status	1 = Member of majority party, 0 = Member of minority party
Legislative Leadership	2 = Major chamber leader, 1 = Standing committee chair, 0 = Rank-and-file
Party Advantage in District	High values indicate support for incumbent's party
Past Election Vote Percentage	Incumbent's vote share in last general election
District Population (thousands)	Number of eligible voters in district divided by number of districts in state
Legislative Professionalism	State legislative professionalism (from Squire (2000))
Chamber Competition	Percentage of seats held by the minority party prior to the election
Chamber (upper house)	1 = Upper house, 0 = Lower house
Year	1 = 1998, 0 = 1996

Original Results

In Table 3, columns 1 and 2, I report the original findings presented in Hogan (2008). The first column reports OLS coefficient estimates and the second includes the conventional OLS standard errors. The main result from this model is the positive and significant coefficient on *Partisan Policy Position.* As Hogan notes, a standard deviation increase in divergence from district

[16] The states represented in the data set are Alaska, California, Florida, Idaho, Illinois, Kentucky, Maine, Michigan, Minnesota, Missouri, Ohio, Oregon, Utah, and Washington.

preferences corresponds to an increase of 5% in the challenger/incumbent spending proportion (2008, 867). On average, as an incumbent's record is increasingly out of step, his or her challenger is able to raise and spend relatively more money.

Table 3: OLS and MR Models Predicting Challenger Spending as a Percentage of Incumbent Spending (Hogan 2008, Table 2)

	OLS			MR		
Variable	β_{OLS}	OLS SE	RCSE	β_{MR}	BSE	KSE
Partisan Policy Position	4.58	1.21	1.02	2.05	1.25	1.27
Political Party	-3.92	2.81	3.69	-3.02	2.44	2.48
Major Party Status	3.93	3.12	5.14	10.86	2.76	2.80
Legislative Leadership	-14.87	3.18	1.79	-10.78	2.91	2.68
Party Advantage in District	-0.21	0.10	0.09	-0.24	0.09	0.09
Past Election Vote Percentage	-0.65	0.09	0.11	-0.90	0.06	0.08
District Population (thousands)	-0.05	0.03	0.06	-0.02	0.01	0.02
Legislative Professionalism	-11.28	1.53	1.98	-10.97	0.98	1.13
Chamber Competition	0.64	0.20	0.25	1.03	0.16	0.19
Chamber (upper house)	-0.99	4.09	6.59	4.69	3.41	3.58
Year	4.07	2.78	2.57	0.69	2.36	2.51
Intercept	98.32	11.28	11.43	80.52	8.06	10.14
Adjusted $R^2/R(p)$	0.14	–	–	0.57	–	–

Note: The dependent variable is challenger spending as a percentage of incumbent spending. See Hogan (2008, 861–865) for a detailed description of the variables included. $R(p)$ is a goodness-of-fit measure for MR; see Koenker and Machado (1999) for more details. The empirical estimate of ρ is 0.044. N = 1,816 clustered in 14 state for both models.

Accounting for Clustering

A key issue with this study that is largely unaddressed is the clustered nature of the data.[17] If the incumbents in the sample are i.i.d., then inference from the model is straightforward. However, it is theoretically reasonable to suspect that incumbents from the same state may be similar in some way. These incumbents are subject to the same rules and legislative norms while in office, the same campaign finance laws during election season, represent citizens in the same geographic area, handle many of the same issues that are unique within states, and may hold similar cultural or political values. As Arceneaux and Nickerson (n.d.) note, there are many potential similarities that exist within states and are distinct across states that are difficult to model.

Despite the theoretical possibility of within-group similarity, few studies in political science attempt to assess the level of clustering in the data by estimating ρ. This problem can easily be remedied, however, because the process is simple to execute in several different software

[17] The issue is not entirely ignored—see page 864, footnote 19.

packages.[18] Using the residuals from the original OLS model reported above, I estimated $\hat{\rho} = 0.044$ for the model reported above. This value falls within the 0.001–0.05 range that is common in medical trials, but it is difficult to say whether it is typical for a political science study. No systematic study has been undertaken on values of ρ within the discipline (Green and Vavreck 2008).[19]

The estimate of 0.044 makes the results comparable to the simulation results presented previously. As a supplement, I performed the simulation procedure under similar conditions to that of the Hogan model ($\rho = 0.044, N = 1,816$, 14 clusters, normal error). Results indicate that the conventional OLS SE were biased downward for the cluster-level variable, but not the individual-level variable (coverage probabilities of 0.81 and 0.95, respectively) and that the RCSE, though still slightly too small, represented an improvement for the cluster-level variable, but not the individual-level variable (coverage probabilities of 0.88 and 0.92, respectively). From this, my next step was to estimate RCSE for the Hogan model to account for the moderate clustering of incumbents within the same state. They are presented in column 3 of Table 3.

A comparison of the two standard error methods highlights some of the findings from the simulations. For instance, there is a difference in how the RCSE affect inferential leverage based on how each variable varies. *Legislative Professionalism*, for instance, varies at the state level (i.e., the level of clustering). The RCSE on that coefficient is almost twice the size of the conventional standard error. As is expected from the simulation results, the RCSE provide better coverage for variables that vary at the cluster level. In contrast, the main independent variable, *Partisan Policy Position*, explains variance only at the individual level, and thus the OLS SE is larger than the RCSE.[20] The common advice from research on standard errors is to report the largest estimate (Green and Vavreck 2008; Arceneaux and Nickerson n.d.). Thus, the OLS SE was the correct choice for the main variable of interest.

[18] In R, ρ can be estimated with the `deff()` command in the `Hmisc` package. In STATA, the command is `loneway`. In SPSS, select the "ICC" option from the reliability analysis menu. SAS users must download one of several user-created macros.

[19] I considered two other replications for this study. A model in Golder (2006) produced $\hat{\rho} = 0.48$ and a model in Brown, Jackson and Wright (1999) produced $\hat{\rho} = 0.78$. Limited simulation work suggests that another technique, such as HLM, may be a better modeling strategy for data clustered at such a high level.

[20] This pattern is broken by some variables. For instance, *Political Party* is an individual-level variable, but the RCSE is larger than the OLS SE. This is likely due to the fact that an incumbent is more likely to be a member of the majority party and the majority party in the legislature can only vary across states. Thus there will be some state-level variation picked up by that variable even though it is measured at the individual level.

MR Replication

In addition to the choice of standard error, it is also important to consider which linear estimator is most appropriate. Using the procedure described in Desmarais and Harden (n.d.), I applied the Vuong (1989) and Clarke (2003, 2007) tests to the Hogan model to determine whether OLS or MR provides a more efficient estimate of the parameters. Both tests selected MR as the more efficient estimator.[21] This suggests that MR is a better method for modeling the average effects of the covariates on challenger spending. Further, the Monte Carlo results reported here demonstrate that within the context of clustered data, the BSE and KSE provide the best coverage of MR coefficient variance. As with the OLS model, I supplemented the results by simulating the MR model under conditions similar to the Hogan data. The coverage probabilities for the individual- and cluster-level variables, respectively, were 0.95 and 0.85 for the BSE and 0.97 and 0.88 for the KSE. The fourth, fifth, and sixth columns of Table 3 report MR coefficients and standard errors for the Hogan model.

The two MR standard error methods show minor differences. As in the Monte Carlo results, the KSE are generally (though not always) larger than the BSE. Following the recommendation of reporting the largest standard error would mean that the KSE should be reported for all of the variables except *Legislative Leadership*. Having selected standard errors, next I consider differences between the OLS and MR results for the Hogan model. Figure 6 plots the coefficient estimates (except the intercept) for both models graphically. The circles appear at the point estimate and the lines extending on either side provide a visual depiction of the 95% confidence intervals. These bounds are computed from the largest standard error for each method reported in Table 3.

---

[21] The Vuong and Clarke tests construct normally- and binomially-distributed test statistics, respectively, from a given sample of data. The tests are based on empirically approximating the ratio of the Kullback-Leibler (KL) divergence of two candidate distributions to a "true" distribution through individual log-likelihood ratios. The null hypothesis in each is that the two models being compared are equidistant from the true model. The procedure described in Desmarais and Harden (n.d.) assumes a normal and Laplace likelihood function for OLS and MR, respectively, and compares the sample data to each. The method corresponding to the distribution that is relatively closer to the sample is deemed the more efficient estimator. In the Hogan model, both tests select MR with p-values less than 0.001. In other words, both tests find that the sample distribution is closer to a Laplace than to a normal, and thus MR is more efficient.

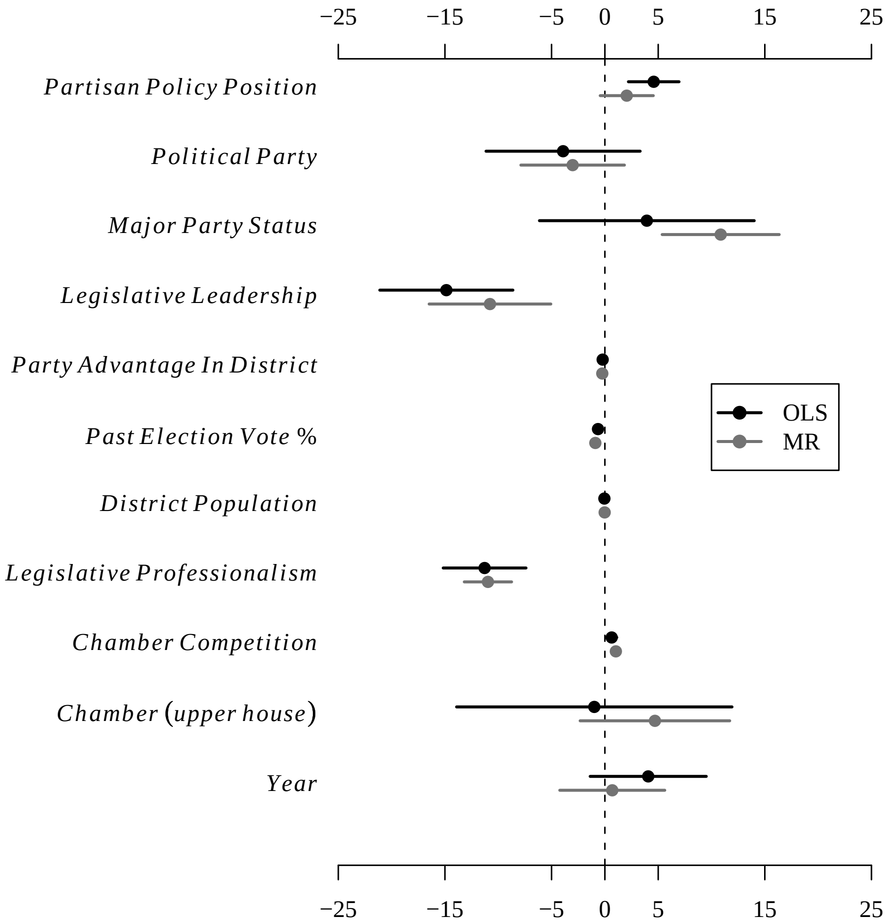

Fig. 6: OLS and MR Models of Hogan (2008, Table 2) with Largest Confidence Interval Estimates

The graph indicates that the general substantive story remains unchanged. With one minor exception (*Chamber (upper house)*), coefficient signs remain the same across the two methods. However, the magnitude of the main independent variable, *Partisan Policy Position*, weakens in the MR model and the 95% confidence interval (computed from the KSE) crosses the vertical line at zero. Thus, according to the more efficient MR model, the effect of *Partisan Policy Position* is not as strong as originally reported.[22]

Conclusions

The results from the simulations and replication reported here indicate several points for political scientists to consider when dealing with clustered data. Overall, they show that the

[22] Hogan reports another OLS model with a different dependent variable (2008, 867). It has roughly the same level of clustering, but substantive results are robust across both OLS and MR specification with all six standard error methods.

standard approach of simply estimating RCSE should not be followed without careful consideration of other factors. While the RCSE approach shows some improvement for statistical inference on a mixed-level data model, the method is not an unconditional remedy for clustering.

One common pattern across all of the simulation results is the benefit of adding clusters. Researchers who have control over the data collection process would be better off adding clusters of data rather than additional observations into old clusters. Adding new clusters corresponds to an increase in the "effective sample size" and reduces the bias caused by the design effect, but adding more observations simply increases the magnitude of the design effect. For instance, a state politics researcher could optimize a data collection process by gathering fewer observations from several different states rather than a large number of observations from only a few states.

Next, although no method performs perfectly, these results support the usefulness of Median Regression for political science research. Two of the standard error methods (BSE and KSE) perform fairly well in clustered data, even when the technique's distributional assumption is violated. In contrast, the OLS methods perform poorly when normality is violated. Thus, this study shows the benefit of estimating the most efficient linear model. The procedure outlined in Desmarais and Harden (n.d.) makes choosing between the two straightforward.

Although political scientists are well aware of clustered data in their own research, almost none of our studies actually measure the degree to which the data are clustered. This study shows that such a measure is crucial for how clustering should be addressed. Estimating RCSE, for instance, really only helps when there is a sufficient level of clustering present. More importantly, the RCSE method actually can have a *negative* impact on statistical inference if it is used when there is little or no clustering. Thus, generating an empirical estimate of ρ should be a standard model diagnostic if clustering is suspected.

Finally, and perhaps most importantly, the level at which each variable in the model varies should be considered. Both the simulations and replication point out that within OLS, the RCSE only help when the variable is explaining cluster-level variance. In other words, there may be a situation in which the OLS SE perform better than the RCSE even when the data are clustered. For each coefficient, the largest standard error estimate should be reported.

Clustered, or mixed-level, data is quite common in political science. This study shows that clustering can cause real issues for statistical inference, but with careful consideration of how many clusters are present in the data, the degree to which the data are clustered, the most efficient regression technique, and the level at which each variable explains variance, the most appropriate standard errors can be selected.

APPENDIX

Computational Information

The Monte Carlo study described here was conducted using R version 2.6.1 (R Development Core Team 2008) with the `quantreg` (Koenker 2008), `Design` (Harrell 2008*a*), `Hmisc` (Harrell 2008*b*), and `mvtnorm` (Genz et al. 2008) packages. Two research computing clusters at the University of North Carolina at Chapel Hill were utilized to carry out the simulations. The first, Emerald, is a general purpose 352-processor Beowulf Linux cluster. Simulations on Emerald were performed using the IBM BladeCenter, Dual Intel Xeon nodes with either 2.4GHz/2.5GB RAM, 2.8GHz/2.5GB RAM, or 3.2GHz/4.0GB RAM. The second, Cedar/Cypress is a 136-processor conifers cluster for scientific applications. Jobs were submitted to the login node (Cedar), which holds 8 Intel Itanium2 processors with 1500MHz/8.0GB RAM and completed by the compute node (Cypress), which holds 128 Intel Itanium2 processors with 1600MHz/512GB RAM (RENCI 2008*a*,*b*).

Coverage Probability

The coverage probability method for a given standard error and coefficient works as follows. The 10,000 estimates of the coefficient (β) are indexed by $j \in (1, 2, \ldots, 9, 999, 10, 000)$. Then a 10,000 $\times$ 3 matrix $\mathbf{O}$ is created with each of the 10,000 coefficient estimates in column 2. Columns 1 and 3 are filled with 95% lower and upper confidence bounds calculated from the standard error estimated for $\beta_{(j)}$. In other words, the standard error is "attached" to its respective coefficient as a confidence interval, as shown below.

$$\mathbf{O} = \begin{bmatrix} \beta_{(1)} - 1.96{\times}\mathrm{SE}_{(1)} & \beta_{(1)} & \beta_{(1)} + 1.96{\times}\mathrm{SE}_{(1)} \\ \beta_{(2)} - 1.96{\times}\mathrm{SE}_{(2)} & \beta_{(2)} & \beta_{(2)} + 1.96{\times}\mathrm{SE}_{(2)} \\ . & . & . \\ . & . & . \\ . & . & . \\ \beta_{(9,999)} - 1.96{\times}\mathrm{SE}_{(9,999)} & \beta_{(9,999)} & \beta_{(9,999)} + 1.96{\times}\mathrm{SE}_{(9,999)} \\ \beta_{(10,000)} - 1.96{\times}\mathrm{SE}_{(10,000)} & \beta_{(10,000)} & \beta_{(10,000)} + 1.96{\times}\mathrm{SE}_{(10,000)} \end{bmatrix}$$

Next, an indicator function $\mathbf{I}(\cdot)$ moves down matrix $\mathbf{O}$, evaluating to one if the confidence interval estimate created in that row includes the true parameter estimate and zero otherwise. A stylized example with five simulations of hypothetical data is show below in which the true value of β is 0.85.

$$\mathbf{SE}^* = \begin{bmatrix} 0.0127 \\ 0.0171 \\ 0.0074 \\ 0.0092 \\ 0.0231 \end{bmatrix} \mathbf{O}^* = \begin{bmatrix} 0.825108 & 0.85 & 0.874892 \\ 0.786484 & 0.82 & 0.853516 \\ 0.775496 & 0.79 & 0.804504 \\ 0.891968 & 0.91 & 0.928032 \\ 0.834724 & 0.88 & 0.925276 \end{bmatrix} \mathbf{I}^* = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Next, the proportion of ones recorded by $\mathbf{I}(\cdot)$ is calculated to produce the coverage probability for that standard error method. In the stylized example, the value $\frac{3}{5} = 0.6$ would be recorded because three of the five simulations produced confidence intervals that covered the true parameter. This process is conducted for each standard error method/coefficient combination, providing a measure with which to compare performance under the changing conditions outlined above. The random number generator seed is set to a common value across all simulations for maximum experimental control and reproducibility.

Sample R code

```
# "Statistical Inference for the Linear Model with Clustered Data"
#
# Jeffrey J. Harden
# University of North Carolina at Chapel Hill
# jjharden@unc.edu
#
# last update: January 29, 2009
#
##########################################################################
# Table of contents:
#  PART I: Loading packages
#  PART II: Creating 95% CI coverage counting function
#  PART III: Monte Carlo simulations
#
# Requires packages quantreg, mvtnorm, Design, and Hmisc
#
# This code is based on the STATA code used in:
#  Green, Donald P. and Lynn Vavreck. 2008. Analysis of Cluster-Randomized Experiments:
#   A Comparison of Alternative Estimation Techniques." Political Analysis 16(2):138-152.
#
##########################################################################
# PART I: Loading packages
library(quantreg)
library(mvtnorm)
library(Design)
library(Hmisc)
#
##########################################################################
# PART II: Creating 95% CI coverage counting function (thanks to Bruce Desmarais)
comp <- function(bse, p){
return(as.numeric(p >= bse[1] - 1.96 * bse[2] & p <= bse[1] + 1.96 * bse[2]))
}

counter <- function(bse, par){
return(mean(apply(bse, 1, comp, p = par)))
}

##########################################################################
# PART III: Monte Carlo simulations

#  Control panel
set.seed(82184)
```

```
n <- 1200 # Sample size
nc <- 50 # Number of clusters
p <- .1
sims <- 10000 # Number of Monte Carlo simulations
boot <- 200 # Number of bootstrap replications
a <- 0
B1 <- 0.85
B2 <- 0.5
cluster <- rep(1:nc, each = n/nc) # Cluster label

OLS.B1 <- numeric(sims) # Vectors to store coefficient and SE estimates
OLS.B2 <- numeric(sims)
OLS.SE1 <- numeric(sims)
OLS.SE2 <- numeric(sims)
OLS.RCSE1 <- numeric(sims)
OLS.RCSE2 <- numeric(sims)
OLS.BSE1 <- numeric(sims)
OLS.BSE2 <- numeric(sims)

MR.B1 <- numeric(sims)
MR.B2 <- numeric(sims)
MR.ASE1 <- numeric(sims)
MR.ASE2 <- numeric(sims)
MR.KSE1 <- numeric(sims)
MR.KSE2 <- numeric(sims)
MR.BSE1 <- numeric(sims)
MR.BSE2 <- numeric(sims)
rho <- numeric(sims)

  for (i in 1:sims){ # Simulate a clustered data set, estimate OLS and MR models

  c.sigma <- matrix(c(4, 0, 0, p), ncol = 2) # Cluster-level random effects
  c.values <- rmvnorm(n = nc, sigma = c.sigma)
  randeff1 <- rep(c.values[ , 1], each = n/nc)
  randeff2 <- rep(c.values[ , 2], each = n/nc)

  i.sigma <- matrix(c(1, 0, 0, (1 - p)), ncol = 2) # Individual-level random effects
  i.values <- rmvnorm(n = n, sigma = i.sigma)
  randeff3 <- i.values[ , 1]
  randeff4 <- i.values[ , 2]

  X1 <- 3 + randeff1 + randeff3 # X1 values unique to individual observations
  X2 <- randeff1 # X2 values unique to clusters of observations
  epsilon <- randeff2 + randeff4 # Two components of the error term

  Y <- a + B1*X1 + B2*X2 + epsilon # True model

  fit.ols <- ols(Y ~ X1 + X2, x = TRUE, y = TRUE) # Model fitting
  fit.mr <- rq(Y ~ X1 + X2)

  OLS.B1[i] <- fit.ols$coef[2]
  OLS.B2[i] <- fit.ols$coef[3]
  OLS.SE1[i] <- sqrt(fit.ols$var[2, 2])
  OLS.SE2[i] <- sqrt(fit.ols$var[3, 3])
  OLS.RCSE1[i] <- sqrt(robcov(fit.ols, cluster)$var[2, 2])
  OLS.RCSE2[i] <- sqrt(robcov(fit.ols, cluster)$var[3, 3])
  OLS.BSE1[i] <- sqrt(bootcov(fit.ols, B = boot)$var[2, 2])
  OLS.BSE2[i] <- sqrt(bootcov(fit.ols, B = boot)$var[3, 3])

  MR.B1[i] <- fit.mr$coef[2]
  MR.B2[i] <- fit.mr$coef[3]
  MR.ASE1[i] <- sqrt((summary(fit.mr, se = "iid", covariance = TRUE)$cov)[2, 2])
  MR.ASE2[i] <- sqrt((summary(fit.mr, se = "iid", covariance = TRUE)$cov)[3, 3])
  MR.KSE1[i] <- sqrt((summary(fit.mr, se = "ker", covariance = TRUE)$cov)[2, 2])
  MR.KSE2[i] <- sqrt((summary(fit.mr, se = "ker", covariance = TRUE)$cov)[3, 3])
  MR.BSE1[i] <- sqrt((summary(fit.mr, se = "boot", R = boot, covariance = TRUE)$cov)[2, 2])
  MR.BSE2[i] <- sqrt((summary(fit.mr, se = "boot", R = boot, covariance = TRUE)$cov)[3, 3])

  rho[i] <- as.numeric(deff(epsilon, cluster)[3])
    if (rho[i] < 0){
      rho[i] <- 0
    }
}
```

```
OLS.cover1 <- counter(cbind(OLS.B1, OLS.SE1), B1) # Coverage probability
OLS.cover2 <- counter(cbind(OLS.B2, OLS.SE2), B2)
OLS.rccover1 <- counter(cbind(OLS.B1, OLS.RCSE1), B1)
OLS.rccover2 <- counter(cbind(OLS.B2, OLS.RCSE2), B2)
OLS.bcover1 <- counter(cbind(OLS.B1, OLS.BSE1), B1)
OLS.bcover2 <- counter(cbind(OLS.B2, OLS.BSE2), B2)

MR.cover1 <- counter(cbind(MR.B1, MR.ASE1), B1)
MR.cover2 <- counter(cbind(MR.B2, MR.ASE2), B2)
MR.kcover1 <- counter(cbind(MR.B1, MR.KSE1), B1)
MR.kcover2 <- counter(cbind(MR.B2, MR.KSE2), B2)
MR.bcover1 <- counter(cbind(MR.B1, MR.BSE1), B1)
MR.bcover2 <- counter(cbind(MR.B2, MR.BSE2), B2)

OLS.mean1 <- mean(OLS.B1) # Coefficient means
OLS.mean2 <- mean(OLS.B2)
MR.mean1 <- mean(MR.B1)
MR.mean2 <- mean(MR.B2)

OLS.meanse1 <- mean(OLS.SE1) # Standard error means
OLS.meanse2 <- mean(OLS.SE2)
OLS.meanrcse1 <- mean(OLS.RCSE1)
OLS.meanrcse2 <- mean(OLS.RCSE2)
OLS.meanbse1 <- mean(OLS.BSE1)
OLS.meanbse2 <- mean(OLS.BSE2)
MR.meanase1 <- mean(MR.ASE1)
MR.meanase2 <- mean(MR.ASE2)
MR.meankse1 <- mean(MR.KSE1)
MR.meankse2 <- mean(MR.KSE2)
MR.meanbse1 <- mean(MR.BSE1)
MR.meanbse2 <- mean(MR.BSE2)

OLS.sd1 <- sd(OLS.B1) # Coefficient standard deviations
OLS.sd2 <- sd(OLS.B2)
MR.sd1 <- sd(MR.B1)
MR.sd2 <- sd(MR.B2)
avgrho <- mean(rho)

results <- cbind(OLS.cover1, OLS.cover2, OLS.rccover1, OLS.rccover2, OLS.bcover1,
OLS.bcover2, MR.cover1, MR.cover2, MR.kcover1, MR.kcover2, MR.bcover1, MR.bcover2,
OLS.mean1, OLS.mean2, MR.mean1, MR.mean2, OLS.meanse1, OLS.meanse2, OLS.meanrcse1,
OLS.meanrcse2, OLS.meanbse1, OLS.meanbse2, MR.meanase1, MR.meanase2, MR.meankse1,
MR.meankse2, MR.meanbse1, MR.meanbse2, OLS.sd1, OLS.sd2, MR.sd1, MR.sd2, avgrho)

write.csv(results, "n1200p1n50.csv", row.names = nc)
```

REFERENCES

Arceneaux, Kevin. 2005. "Using Cluster-Randomized Field Experiments to Study Voting Behavior." *Annals of the American Academy of Political and Social Science* 601(1):169–179.

Arceneaux, Kevin and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53(1):1–16.

Arceneaux, Kevin and David W. Nickerson. n.d. "Modeling Certainty with Clustered Data: A Comparison of Methods." Unpublished manuscript.

Arceneaux, Kevin and Gregory Huber. 2007. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science* 51(4):957–977.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics* 49(4):431–434.

Barreto, Matt A., Gary M. Segura and Nathan D. Woods. 2004. "The Mobilizing Effect of Majority-Minority Districts on Latino Turnout." *American Political Science Review* 98(1):65–75.

Bassett, Gilbert and Roger Koenker. 1978. "Asymptotic Theory of Least Absolute Error Regression." *Journal of the American Statistical Association* 73(363):618–622.

Beck, Nathaniel and Jonathan N. Katz. 1995. "What to Do (And Not to Do) With Time-Series Cross-Section Data." *American Political Science Review* 89(3):634–647.

Berry, Christopher. 2008. "Multilevel Government and the Fiscal Common-Pool." *American Journal of Political Science* 52(4):802–820.

Bonneau, Chris W. 2005. "What Price Justice(s)? Understanding Campaign Spending in State Supreme Court Elections." *State Politics and Policy Quarterly* 5(2):107–125.

Bradlow, Eric T., Howard Wainer and Xiaohui Wang. 1999. "A Bayesian Random Effects Model for Testlets." *Psychometrika* 64(2):153–168.

Branton, Regina P. 2004. "Voting in Initiative Elections: Does the Context of Racial and Ethnic Diversity Matter?" *State Politics and Policy Quarterly* 4(3):294–317.

Brown, Robert D., Robert A. Jackson and Gerald C. Wright. 1999. "Registration, Turnout, and State Party Systems." *Political Research Quarterly* 52(3):463–479.

Buckley, Jack and Chad Westerland. 2004. "Duration Dependence, Functional Form, and Corrected Standard Errors: Improving EHA Models of State Policy Diffusion." *State Politics and Policy Quarterly* 4(1):94–113.

Büthe, Tim and Helen V. Milner. 2008. "The Politics of Foreign Direct Investment into Developing Countries: Increasing FDI through International Trade Agreements?" *American Journal of Political Science* 52(4):741–762.

Carsey, Thomas M. and Gerald C. Wright. 1998. "State and National Factors in Gubernatorial and Senatorial Elections." *American Journal of Political Science* 42(3):994–1002.

Carsey, Thomas M. and Robert A. Jackson. 2001. "Misreport of Vote Choice in U.S. Senate and Gubernatorial Elections." *State Politics and Policy Quarterly* 1(2):196–209.

Carson, Jamie L. and Michael H. Crespin. 2004. "The Effect of State Redistricting Methods on Electoral Competition in United States House of Representatives Races." *State Politics and Policy Quarterly* 4(4):455–469.

Clarke, Kevin A. 2003. "Nonparametric Model Discrimination in International Relations." *Journal of Conflict Resolution* 47(1):72–93.

Clarke, Kevin A. 2007. "A Simple Distribution-Free Test for Nonnested Hypotheses." *Political Analysis* 15(3):347–363.

Cornfield, Jerome. 1978. "Randomization by Group: A Formal Analysis." *American Journal of Epidemiology* 108(2):100–102.

Crescenzi, Mark J. C. 2007. "Reputation and Interstate Conflict." *American Journal of Political Science* 51(2):382–396.

Danilovic, Vesna and Joe Clare. 2007. "The Kantian Liberal Peace (Revisited)." *American Journal of Political Science* 51(2):397–414.

Desmarais, Bruce A. and Jeffrey J. Harden. n.d. "Efficient Estimation of the Linear Model: Choosing Between Conditional-Mean and Conditional-Median Methods." Unpublished manuscript.

Donner, Allan and George Wells. 1986. "A Comparison of Confidence Interval Methods for the Intraclass Correlation Coefficient." *Biometrics* 42(2):401–412.

Donner, Allan and Neil Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. New York: Arnold Publishers.

Donner, Allan and Neil Klar. 2004. "Pitfalls of and Controversies in Cluster Randomization Trials." *American Journal of Public Health* 94(3):416–422.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7(1):1–26.

Efron, Bradley. 1982. "Nonparametric Estimates of Standard Error: the Jackknife, the Bootstrap and Other Methods." *Biometrika* 68(3):589–599.

Gabel, Matthew and Kenneth Scheve. 2007. "Estimating the Effect of Elite Communications on Public Opinion Using Instrumental Variables." *American Journal of Political Science* 51(4):1013–1028.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge, MA: Cambridge University Press.

Genz, Alan, Frank Bretz, Torsten Hothorn with contributions by Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch and Fabian Scheipl. 2008. *mvtnorm: Multivariate Normal and t Distributions.* R package version 0.9–3.
URL: *http://CRAN.R-project.org/package=mvtnorm*

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102(1):33–48.

Golder, Matt. 2006. "Presidential Coattails and Legislative Fragmentation." *American Journal of Political Science* 50(1):34–48.

Gould, William. 1992. "Quantile Regression With Bootstrapped Standard Errors." *Stata Technical Bulletin* pp. 19–21.

Green, Donald P. and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Techniques." *Political Analysis* 16(2):138–152.

Greene, William H. 2002. *Econometric Analysis.* 5 ed. Upper Saddle River, NJ: Prentice Hall.

Gujarati, Damodar N. 2003. *Basic Econometrics.* International edition ed. New York: McGraw-Hill.

Hao, Lingxin and Daniel Q. Naiman. 2007. *Quantile Regression.* Quantitive Applications in the Social Sciences Thousand Oaks, CA: Sage Publications.

Harrell, Frank E. 2008*a*. *Design: Design Package.* R package version 2.1–2.
URL: *http://biostat.mc.vanderbilt.edu/s/Design*

Harrell, Frank E. 2008*b*. *Hmisc: Harrell Miscellaneous.* R package version 3.4–4.
URL: *http://biostat.mc.vanderbilt.edu/s/Hmisc*

Hogan, Robert E. 2008. "Policy Responsiveness and Incumbent Reelection in State Legislatures." *American Journal of Political Science* 52(4):858–873.

Huber, P.J. 1967. The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press pp. 221–233.

Kasara, Kimuli. 2007. "Tax Me If You Can: Ethnic Geography, Democracy, and the Taxation of Agriculture in Africa." *American Political Science Review* 101(1):159–172.

Killip, Shersten, Ziyad Mahfoud and Kevin Pearce. 2004. "What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers." *The Annals of Family Medicine* 2(3):204–208.

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

Koenker, Roger. 1994. Confidence Intervals for Regression Quantiles. In *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*. New York: Springer-Verlag pp. 349–359.

Koenker, Roger. 2005. *Quantile Regression*. Cambridge, UK: Cambridge University Press.

Koenker, Roger. 2008. *quantreg: Quantile Regression*. R package version 4.24.
URL: *http://www.r-project.org/*

Koenker, Roger and Gilbert Bassett. 1978. "Regression Quantiles." *Econometrica* 46(1):33–50.

Koenker, Roger and José A. F. Machado. 1999. "Goodness of Fit and Related Inference Processes for Quantile Regression." *Journal of the American Statistical Association* 94(448):1296–1310.

Koenker, Roger and Kevin F. Hallock. 2001. "Quantile Regression: An Introduction." *Journal of Economic Perspectives* 15(4):143–156.

Liu, Regina Y. 1988. "Bootstrap Procedure under some non-I.I.D. models." *The Annals of Statistics* 16(4):1696–1708.

Long, J. Scott and Laurie H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *The American Statistician* 54(3):217–224.

McDonald, Patrick J. 2007. "The Purse Strings of Peace." *American Journal of Political Science* 51(3):568–582.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics* 72(2):334–338.

Newcombe, Robert G. 1998. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods." *Statistics in Medicine* 17(8):857–872.

Platt, Robert W., James A. Hanley and Hong Yang. 2000. "Bootstrap Confidence Intervals for the Sensitivity of a Quantitative Diagnostic Test." *Statistics in Medicine* 19(3):313–322.

Posner, Daniel N. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98(4):529–545.

Powell, James L. 1991. Estimation of Monotonic Regression Models Under Quantile Restrictions. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. William A. Barnett, James L. Powell and George Tauchen. Cambridge: Cambridge University Press.

Primo, David M., Matthew L. Jacobsmeier and Jeffrey Milyo. 2007. "Estimating the Impact of State Policies and Institutions with Mixed-Level Data." *State Politics and Policy Quarterly* 7(4):446–459.

R Development Core Team. 2008. *R : A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
URL: *http://www.r-project.org*

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* 2nd ed. Thousand Oaks, CA: Sage Publications.

RENCI. 2008*a*. *Cedar/Cypress.* Chapel Hill, NC: Renaissance Computing Institute, University of North Carolina at Chapel Hill. Accessed December 11, 2008.
URL: *http://www.renci.org/unc/computing/resources/cedcyp.php*

RENCI. 2008*b*. *Emerald.* Chapel Hill, NC: Renaissance Computing Institute, University of North Carolina at Chapel Hill. Accessed December 11, 2008.
URL: *http://www.renci.org/unc/computing/resources/emerald.php*

Rogers, William. 1992. "Quantile Regression Standard Errors." *Stata Technical Bulletin* pp. 16–19.

Squire, Peverill. 2000. "Uncontested Seats in State Legislative Elections." *Legislative Studies Quarterly* 25(1):131–146.

Steenbergen, Marco R. and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46(1):218–237.

Stratmann, Thomas. 2006. "Is Spending More Potent For or Against a Proposition? Evidence from Ballot Measures." *American Journal of Political Science* 50(3):788–801.

Tolbert, Caroline J., Ramona S. McNeal and Daniel A. Smith. 2003. "Enhancing Civic Engagement: The Effect of Direct Democracy on Political Participation and Knowledge." *State Politics and Policy Quarterly* 3(1):23–41.

Ukoumunne, Obioha C. 2002. "A Comparison of Confidence Interval Methods for the Intraclass Correlation Coeffcient in Cluster Randomized Trials." *Statistics in Medicine* 21(24):3757–3774.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2):307–333.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–838.

Williams, Rick L. 2000. "A Note on Robust Variance Estimation for Cluster-Correlated Data." *Biometrics* 56(2):645–646.

Wolfinger, Raymond E., Benjamin Highton and Megan Mullin. 2005. "How Postregistration Laws Affect the Turnout of Citizens Registered to Vote." *State Politics and Policy Quarterly* 5(1):1–23.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93(2):133–138.

Yu, Keming and Rana Moyeed. 2001. "Bayesian Quantile Regression." *Statistics and Probability Letters* 54(4):437–447.

Zeger, Scott L. and Kung-Yee Liang. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics* 42(1):121–130.

Zorn, Christopher. 2006. "Comparing GEE and Robust Standard Errors for Conditionally Dependent Data." *Political Research Quarterly* 59(3):329–341.