State Politics & Policy Quarterly
http://spa.sagepub.com/

A Bootstrap Method for Conducting Statistical Inference with Clustered Data

Jeffrey J. Harden

State Politics & Policy Quarterly 2011 11: 223
DOI: 10.1177/1532440011406233

The online version of this article can be found at:
http://spa.sagepub.com/content/11/2/223

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
American Political Science Association

State Politics & Policy Quarterly
11(2) 223–246

DOI: 10.1177/1532440011406233
http://sppq.sagepub.com
SAGE

A Bootstrap Method for Conducting Statistical Inference with Clustered Data

Jeffrey J. Harden[1]

Abstract

U.S. state politics researchers often analyze data with observations grouped into clusters. This structure commonly produces unmodeled correlation within clusters, leading to downward bias in the standard errors of regression coefficients. Estimating robust cluster standard errors (RCSE) is a common approach to correcting this bias. However, despite their frequent use, recent work indicates that RCSE can also be biased downward. Here the author provides evidence of that bias and offers a potential solution. Through Monte Carlo simulation of an ordinary least squares (OLS) regression model, the author compares conventional standard error (OLS-SE) and RCSE performance to that of a bootstrap method that resamples clusters of observations (BCSE). The author shows that both OLS-SE *and* RCSE are biased downward, with OLS-SE being the most biased. In contrast, BCSE are not biased and consistently outperform the other two methods. The author concludes with three replications from recent work and offers recommendations to researchers.

Keywords

clustered data, standard errors, bootstrapping

One problem that political scientists must often confront when analyzing data is the possibility of unmodeled correlation between observations within groups, commonly referred to as "clustered" or "multilevel" data. U.S. state politics scholars face this issue with observations grouped in different states, legislative districts, or

[1]University of North Carolina at Chapel Hill, USA

Corresponding Author:
Jeffrey J. Harden, University of North Carolina at Chapel Hill, Department of Political Science, 312 Hamilton Hall, CB #3265, Chapel Hill, NC 27599
Email: jjharden@unc.edu

counties. Clustering may also occur because of repeated observations from the same actor, such as several votes by a judge or legislator.[1] Researchers should be concerned with this unique data structure because it can create downward bias in the standard errors of regression coefficients, leading to a higher likelihood of committing a Type I error—rejecting the null hypothesis when it is actually true (Arceneaux 2005; Arceneaux and Nickerson 2009; Green and Vavreck 2008; Moulton 1990).

One common approach to solving this problem is to account for clustering by estimating robust cluster standard errors (RCSE). However, despite this frequent use, scholars often merely assume that RCSE are providing an accurate picture of coefficient variability and rarely investigate this assumption. In political science, only a few recent studies assess their performance in systematic fashion. This work demonstrates conditions under which RCSE are themselves biased downward and shows that they are not always the best option for handling clustered data (Arceneaux and Nickerson 2009; Cameron, Gelbach, and Miller 2008; Green and Vavreck 2008).

In this study, I examine an alternative to RCSE: a bootstrap method that I refer to as "bootstrap cluster standard errors" (BCSE).[2] The method, which can be used with linear models (e.g., ordinary least squares [OLS] or one of its variants) or a wide variety of generalized linear models (e.g., logit, probit, or duration models), addresses the clustered structure of the data nonparametrically by resampling entire clusters of observations (with replacement) to calculate standard errors. Past work shows that this approach provides more accurate estimates of coefficient variability in clustered data than RCSE (Cameron, Gelbach, and Miller 2008). I show similar results here and expand on those findings by demonstrating that BCSE provide the most improvement for variables exhibiting cluster-level variation and that RCSE can still be too small with large numbers of clusters.

My first means of assessment is to compare conventional OLS standard errors (OLS-SE), RCSE, and BCSE via Monte Carlo simulation. Then I replicate three recent analyses to show evidence of external validity of the simulation results and to demonstrate that BCSE can have important substantive implications for inference from statistical models in state politics. More specifically, I report the following findings:

1. OLS-SE are biased downward in clustered data;
2. RCSE perform better than OLS-SE but are also biased downward under a range of simulated conditions;
3. BCSE consistently provide more accurate estimates of coefficient variability in clustered data than OLS-SE and RCSE;
4. The differences among the three methods are smaller for variables that exhibit individual-level variation and larger for variables that exhibit cluster-level variation;
5. Estimation with BCSE can yield different substantive conclusions from statistical models than OLS-SE and RCSE in state politics research;

From these results I recommend that state politics researchers use BCSE to conduct statistical inference with clustered data. In the online appendix (available at http://

academic.udayton.edu/SPPQ-TPR/index.htm), I offer advice on implementing BCSE in applied work and provide example code for using the method in R and Stata.

The Problem of Clustering

Though clustered data are not unique to the discipline, the structure is common in political science. Within the study of U.S. state politics, researchers are especially likely to encounter observations grouped together in some way (e.g., Arceneaux and Huber 2007; Brown, Jackson, and Wright 1999; Carsey and Jackson 2001; Hogan 2008; Tolbert, McNeal, and Smith 2003; Wolfinger, Highton, and Mullin 2005). Furthermore, additional research indicates that not accounting for the clustered nature of the data can pose problems for statistical inference even when there is no omitted variable bias (Arceneaux and Nickerson 2009; Carsey and Wright 1998; Green and Vavreck 2008; Primo, Jacobsmeier, and Milyo 2007; Zorn 2006). This is the result of the fact that a single observation in a clustered data set contributes less unique information to the model than in data where no clustering exists. This is formalized by what Kish (1965) calls the design effect (DEFF).

The Design Effect

Clustering in the data constitutes a violation of the assumption of independent errors. This means that the "effective sample size" is not the total number of observations but rather closer to the number of clusters in the data (Arceneaux and Nickerson 2009). Formally, the measure of similarity between observations within clusters is the intracluster correlation coefficient, or ρ. This value is defined as,

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

where $s^2_{between}$ is the variance of the error term across cluster means and s^2_{within} is the mean variance within clusters. The DEFF is a measure of the bias to conventional statistical inference methods in a given sample of data, and is defined as,

$$\text{DEFF} = 1 + \rho \,.\, (N/C - 1) \tag{2}$$

where N is the sample size and C is the number of clusters (Kish 1965).[3] From this, conventional standard errors will be biased downward by a factor of $\sqrt{\text{DEFF.}}$[4]

Approaches to Dealing with Clustered Data

To this point, most research on clustered data in political science has assumed that the RCSE method adequately accounts for the DEFF by reporting results in which the RCSE estimates are larger than the OLS-SE estimates (e.g., Arceneaux 2005; Zorn 2006; but see Arceneaux and Nickerson 2009; Green and Vavreck 2008). However,

work in economics and statistics finds RCSE to be biased downward, particularly with small numbers of clusters (see Cameron, Gelbach, and Miller 2008).[5] In light of this, it may be more beneficial to use another estimation strategy to deal with clustered data, such as cluster-level randomization tests (CLRT; Erikson, Pinto, and Rader 2010), aggregation to the cluster level (Green and Vavreck 2008), estimation by generalized estimating equation (GEE) models (Liang and Zeger 1986; Zorn 2001), or multilevel modeling (MLM; Gelman and Hill 2007; Raudenbush and Bryk 2002). However, all of these methods have their own benefits and limitations.

MLM, for instance, accounts for the unobserved cluster effects explicitly through the likelihood function but can be more demanding of the data because it estimates more parameters. This can lead to nonconvergence in some models, such as those with many cross-level interactions and/or those estimated with small samples of data (e.g., Primo, Jacobsmeier, and Milyo 2007). GEE does not require parametric specification of the errors, but interpretation of results is slightly different and the method does not have nearly the same goodness-of-fit assessment capabilities as likelihood-based models (Zorn 2001).[6] Aggregation to the cluster level solves the issue of cluster correlation but also eliminates the possibility of specifications with cross-level interactions, which are often theoretically interesting to scholars of state politics (e.g., Wolfinger, Highton, and Mullin 2005).

Finally, CLRT, which conducts hypothesis testing by reshuffling the values of cluster-level variables, is similar to the BCSE method shown below. Erikson, Pinto, and Rader (2010) provide evidence that CLRT is effective because of its nonparametric nature. However, CLRT assumes exchangeability of errors (Erikson, Pinto, and Rader 2010; Kennedy 1995), leads to a different (though not incorrect) conceptualization of model error (Kennedy 1995), and cannot calculate covariance between coefficient estimates. Bootstrapping by clusters does not make the exchangeability assumption, follows the same conception of model error that most political scientists are familiar with, and estimates the full covariance matrix.[7] See the appendix for more details.

Standard Error Methods

In a typical linear model such as $y = \mathbf{X}\beta+\varepsilon$ in which $\mathbf{X}$ is a matrix of independent variables with a leading vector of ones and $\varepsilon \sim N(0,\sigma)$, the OLS covariance matrix of the slope coefficients is calculated as,

$$\mathrm{OLS}_{\mathrm{cov}}(\hat{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{X}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \tag{3}$$

where $\boldsymbol{\Phi} = (\hat{\varepsilon}\hat{\varepsilon}^{\mathrm{T}})$. If ε is homoscedastic and $\mathrm{cov}(\varepsilon_i,\varepsilon_j = 0)\ \forall\ i \neq j$, then $\boldsymbol{\Phi} = \sigma^2\mathbf{I}$ and the right-hand side of equation 3 reduces to $\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. OLS-SE can then be calculated as the square roots of the elements on the main diagonal of this covariance matrix. Of course, the assumption that the errors are independent is violated when clustering is present.

RCSE

To account for this, RCSE constitute a simple adjustment to equation 3 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 zero as they are in the conventional calculation. The robust cluster covariance matrix calculation takes the form,

$$\mathrm{RC}_{\mathrm{cov}}(\hat{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sum_{c=1}^{c}\{(\sum_{i=1}^{n_c}\hat{\varepsilon}_{ic}\mathbf{x}_{ic})(\sum_{i=1}^{n_c}\hat{\varepsilon}_{ic}x_{ic})^{\mathrm{T}}\}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \quad (4)$$

where n_c is the number of observations in cluster c, ε_{ic} is the residual for the i^{th}observation of cluster c, and $\mathbf{x}_{ic}$ is a row vector of independent variables for the same observation. In addition, for a model with K regressors, most software packages apply the finite sample adjustment $C/C{-}1 \cdot N{-}1/N{-}K$ (Primo, Jacobsmeier, and Milyo 2007, 451).

This is an example of a "sandwich estimator" like the method behind the well-known Huber–White heteroscedasticity-robust standard errors (HWSE; see Huber 1967; White 1980).[8] The name refers to the mathematical form of equation 4, which looks like a sandwich with the "bread" as the familiar $(X^{\mathrm{T}}X)^{-1}$. The "meat" of the sandwich is the outer product of cluster-level scores, calculated by summing individual-level score vectors within each cluster (Franzese 2005; Rogers 1993). By incorporating nonzero covariances within clusters, RCSE are designed to be robust to heteroscedasticity *and* cluster correlation, though observations from different clusters are still assumed independent (Williams 2000).

BCSE

An alternative to the parametric assumptions necessary for OLS-SE and RCSE is the nonparametric bootstrap. Bootstrapping is a general method that allows for the approximation of the sampling distribution of a statistic via simulation from the observed data (Efron and Tibshirani 1993). The method operates as follows. Denote a sample $S = \{x1,x2,x3,\ldots,x_N\}$ of size N drawn from a population P. A statistic of interest, Θ, that describes P can be estimated by calculating $\hat{\Theta}$ from S. Then the sampling variability of $\hat{\Theta}$ can be calculated via the bootstrap in the following way:

1. Draw a sample of size N from S with replacement such that each element is selected with probability 1/N. Denote this a "bootstrap sample," S_{boot}.
2. Calculate a new estimate of Θ from S_{boot}. Denote this bootstrap estimate $\hat{\Theta}^*_j$.
3. Repeat steps 1 and 2 B times, storing each $\hat{\Theta}^*_j$ to create $\mathbf{V}$, a vector of bootstrap estimates.

For a sufficiently large B, the standard deviation of $\mathbf{V}$ can then be treated as the standard error of $\hat{\Theta}$, which can be used to construct confidence intervals for the estimate.

Alternatively, the quantiles of **V** could be used to form a completely nonparametric confidence interval (e.g., 0.025 and 0.975 for a 95% confidence interval).

The BCSE method, which is based on a time-series application due to Künsch (1989), tailors the bootstrap logic of resampling with replacement to the structure of clustered data (Cameron, Gelbach, and Miller 2008; Cameron and Trivedi 2005; Feng, McLerran, and Grizzle 1996). The only modification is in step 1. Instead of drawing a sample of size N from the individual observations in S, the method draws a sample of C clusters of observations from S with replacement such that each cluster is selected with equal probability.

As a result of this process, the individual observations within a cluster all "move" into the bootstrap samples together, reflecting the true form of the dependencies between observations. In addition, the BCSE method requires only the assumption that observations from different clusters are independent. As mentioned above, this sets it apart from CLRT, which requires the independence assumption *and* the assumption that errors are exchangeable (Erikson, Pinto, and Rader 2010; Kennedy 1995).[9]

Monte Carlo Simulations

In this section I provide a comparison of the three methods detailed above using Monte Carlo simulation. In particular, I systematically vary all three parameters that influence the DEFF: number of clusters (C), sample size (N), and intracluster correlation (ρ). In addition, I expand on past simulation work that looks primarily at individual-level variables (e.g., Cameron, Gelbach, and Miller 2008; Green and Vavreck 2008) and examine standard error performance for variables that exhibit both individual- and cluster-level variation.[10]

First, I simulate data with 10, 25, 40, 50, and 100 clusters. Previous work indicates that adding clusters decreases conventional standard error bias (e.g., Arceneaux and Nickerson 2009; Green and Vavreck 2008). In other words, for a constant value of N, OLS-SE should become less biased as the number of clusters increases. RCSE are also known to perform better with more clusters (Cameron and Trivedi 2005, 834). The maximum value of 100 is chosen to reflect the number of clusters most state politics researchers can expect to find in their data, but results (not shown) are consistent at higher values, such as 200 clusters.[11]

Next, I consider changes to the sample size by conducting the simulation with 200, 800, and 1,200 observations. These values are selected to reflect realistic sample sizes in the study of state politics. A larger sample size should produce OLS-SE estimates that are increasingly too small. Consider the formula for the DEFF (equation 2); holding ρ constant, the DEFF will increase as N increases. This, in turn, should increase the magnitude of the downward bias to OLS-SE. By adding more observations to the same number of clusters, the clustered structure in the data becomes more pronounced, making a conventional inference method that assumes independent errors more problematic.

Finally, I conduct the simulations at two values of ρ: 0.10, which is the value used in the Green and Vavreck (2008) study, and 0.50.[12] Increasing the value of ρ will increase the DEFF, which should cause OLS-SE to be increasingly biased downward.

The more important question is whether RCSE and BCSE can accommodate the stronger dependencies within clusters.

Model Estimation

To assess the effects of these parameters I simulate the following linear 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 $i \in (1,2,3,\ldots,N)$ and c clusters of observations $c \in (1,2,3,\ldots,C)$ I also include a multiplicative interaction between the two independent variables. This is done to create variation at three different levels: x_1 varies only at the individual level, x_2 varies only at the cluster level, and the interaction exhibits both between-cluster variation *and* variation within clusters.

$$y_{ic} = \alpha + \beta_1 x_{1ic} + \beta_2 x_{2c} + \beta_3 (x_{1ic} \cdot x_{2c}) + \varepsilon_{ic} \quad (5)$$

The model is defined such that $\alpha = 0$, $\beta = .85$, $\beta = .50$, and $\beta = .70$, although these values can be changed without affecting results. In addition, the error term (ε_{ic}), which has a mean of zero and is uncorrelated with the independent variables, is broken into two components: an individual-level disturbance e_{ic} and a cluster-level disturbance v_c (both distributed normally) such that $\varepsilon_{ic} = e_{ic} + v_c$.

To assess standard error performance I calculate coverage probabilities. This process involves constructing 95% confidence intervals from the standard errors produced by each simulation and calculating the proportion that includes the true parameter. The expectation is that this proportion should be 0.95 if the standard error method is "correct."[13] A value less than 0.95 indicates downward bias (i.e., toward Type I errors) and a value greater than 0.95 indicates a conservative (Type II error) bias.[14]

Results

The simulation procedure is composed of 10,000 replications at the five different values of C (10, 25, 40, 50, and 100), the three different values of N (200, 800, and 1,200), and each value of ρ (0.10 and 0.50). This yields a total of 300,000 simulations.[15] I present selected results graphically, with the number of clusters plotted across the x-axes and the coverage probabilities of each method on the y-axes.[16] The corresponding DEFF is labeled on each x-axis to allow for comparability across graphs. In addition, a line is drawn at 0.95 to indicate the standard for what a 95% confidence interval should cover, with dashed lines indicating 95% Monte Carlo confidence bounds to account for simulation error. A standard error estimate is unbiased if its coverage probability falls within those dashed lines.[17]

Number of Clusters and Sample Size

I begin with the implications of changing C and N. The top panels of Figure 1 plot the coverage probabilities for the coefficient on the individual-level variable (β_1) at sample

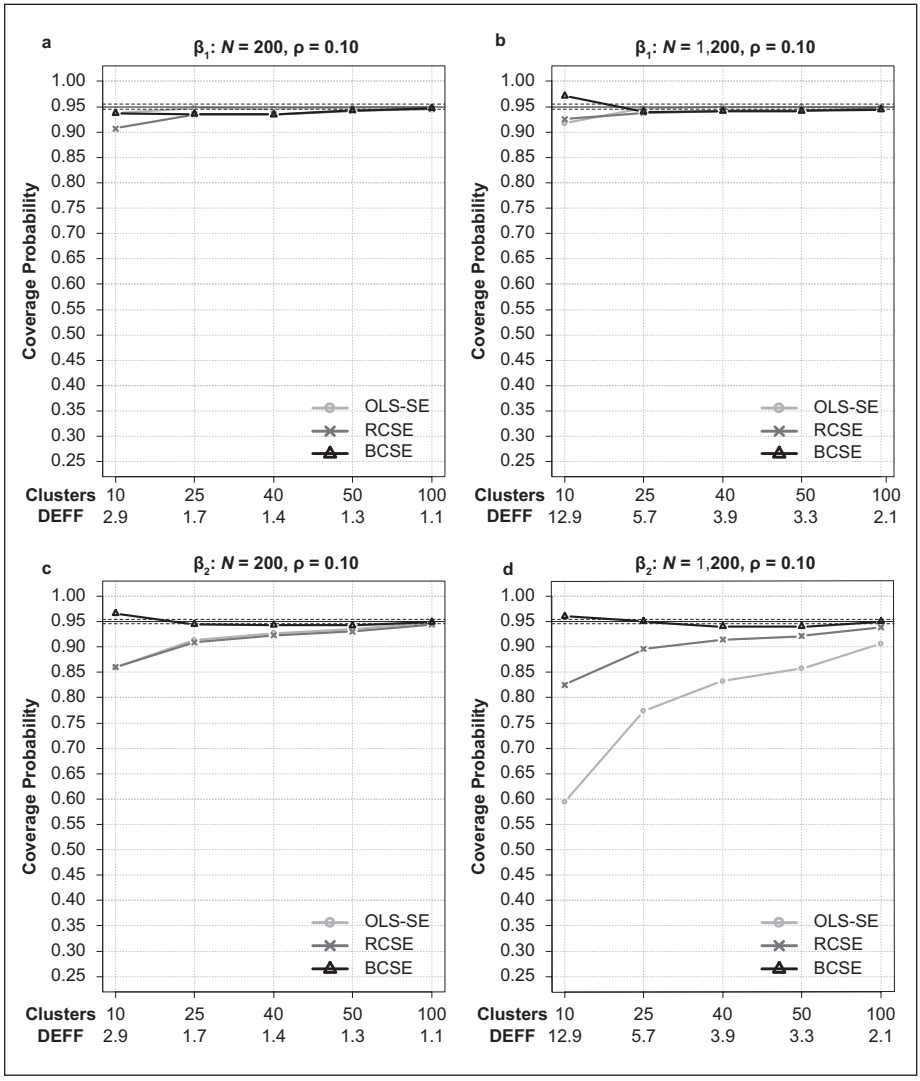

Figure 1. Effects of increasing *C* and *N* for β_1 and β_2 with $\rho = 0.10$

Note: The graphs plot coverage probabilities for β_1 (panels a and b) and β_2 (panels c and d) with $N = 200$ and 1,200 and $\rho = 0.10$. The black dashed lines represent 95% Monte Carlo simulation error bounds.

sizes of 200 (panel a) and 1,200 (panel b) and ρ held constant at 0.10. The bottom panels plot the same conditions for the cluster-level variable (β_2). Within each graph the DEFF decreases across the x-axes, but it is larger in absolute terms when $N = 1{,}200$.

The first point to note is that the standard errors for the coefficient on the individual-level variable (β_1) are essentially unaffected by the clustering—all three methods are close to 0.95 across the range of clusters in the top panels of Figure 1. Adding a cluster-level effect to the errors has almost no impact when the variation of a variable is entirely at the observation level.

However, that is not the case when a variable exhibits between-cluster variation. The bottom panels of Figure 1 show downward bias in the coverage probabilities of OLS-SE and RCSE estimated for β_2. An increase in the number of clusters decreases this bias to the OLS-SE and RCSE estimates, as shown by the upward-trending lines in those graphs. However, increasing the sample size corresponds with an increase in the downward bias to the two methods. At $N = 200$ and $C = 10$ (panel c), the OLS-SE and RCSE coverage probabilities are each 0.86 (the two methods are nearly identical), while at $N = 1{,}200$ and $C = 10$ (panel d) the coverage probabilities are 0.59 (OLS-SE) and 0.83 (RCSE). In contrast, BCSE estimated for β_2 do not exhibit this variation. They consistently fall closer to 0.95 for all five cluster values. In some cases they barely miss the simulation error bounds—both upper and lower—but overall they perform much better than OLS-SE and RCSE. Similar results hold for β_3 (see the full results online).

Intracluster Correlation

The other parameter that affects the magnitude of the DEFF is the level of correlation within clusters. Figure 2 shows the coverage probability results for the coefficient on the cluster-level variable (β_2) and the cross-level interaction (β_3) at the two different values of ρ and a constant sample size of 800. Again the DEFF decreases across the x-axis within each graph, but in this case it increases absolutely as ρ increases from the left panels to the right panels.

The results show that OLS-SE become increasingly biased downward as ρ increases. When $\rho = 0.10$ (panels a and c), the highest OLS-SE coverage probability is 0.92 for β_2 and 0.88 for β_3 at 100 clusters, but these values drop to 0.81 and 0.66, respectively, when $\rho = 0.50$ (panels b and d). RCSE are also biased downward but show a slightly different pattern; they perform at about the same levels at both values of ρ, with coverage probabilities between 0.79 and 0.94 in all four panels. However, despite this consistency, RCSE are still (slightly) biased downward even at 100 clusters in Figure 2. Most importantly, BCSE consistently perform well. Aside from being slightly too conservative at 10 clusters, they stay close to coverage probabilities of 0.95 across the range of clusters at both levels of intracluster correlation.

Monte Carlo Summary

Overall, several results emerge from these simulations. First, OLS-SE and RCSE are both biased downward under a wide range of conditions. In the case of OLS-SE, this bias is the result of the incorrect assumption that the observations are independent. However, the source of the bias to RCSE is less obvious because there is proof that

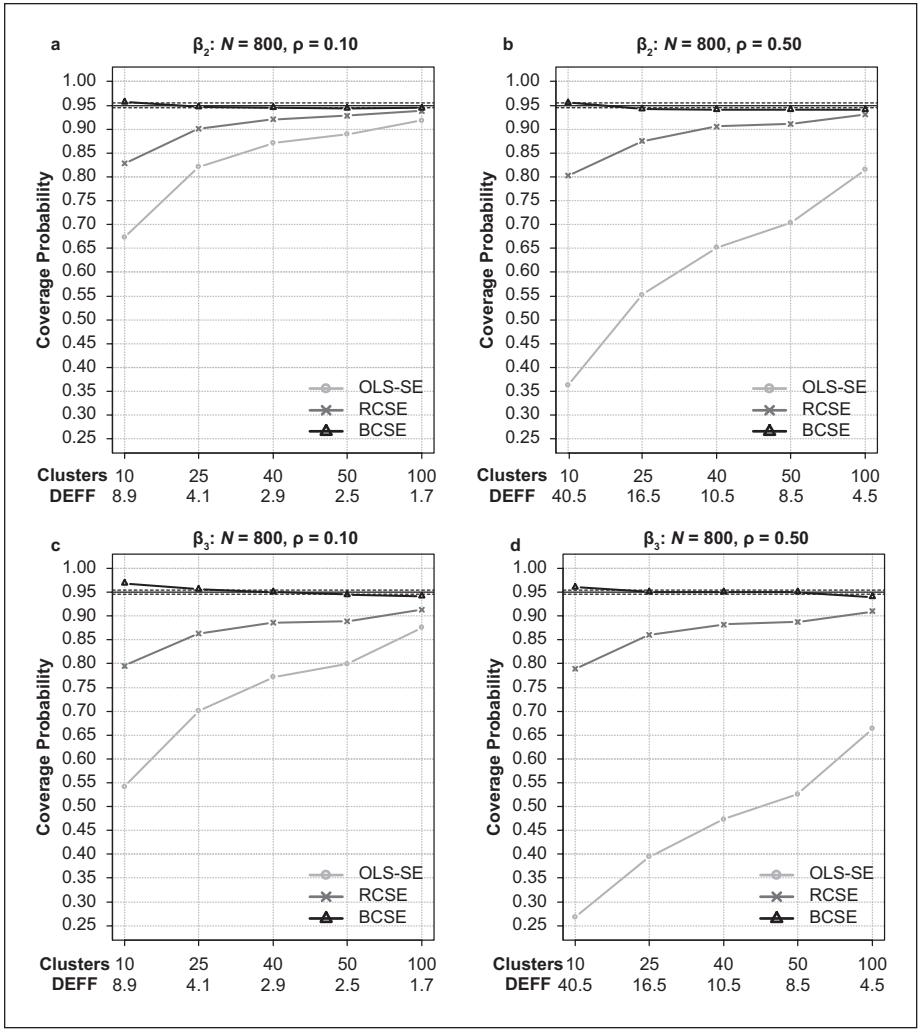

Figure 2. Effects of increasing ρ for β_2 and β_3 with $N = 800$

Note: The graphs plot coverage probabilities for β_2 (panels a and b) and β_3 (panels c and d) with $N = 800$ and $\rho = 0.10$ and 0.50. The black dashed lines represent 95% Monte Carlo simulation error bounds.

the method is an unbiased estimator (Williams 2000). Cameron, Gelbach, and Miller (2008, 415) note that the RCSE method suffers because it relies on the model residuals to estimate the cluster correlation in the data and those residuals are biased estimates of the true error in the model. My own simulations support this; see the full results online for details.[18]

Second, a novel contribution from these simulations is the demonstration of the poor performance of RCSE even at large numbers of clusters for variables that exhibit cluster-level variation (i.e., Figure 2). Past simulation work focuses more on individual-level variables, where RCSE can still be biased but improve as more clusters are added. Here I show that RCSE are still biased at 100 clusters if the variable exhibits between-cluster variation.

Finally, the most important finding is that BCSE performance remains consistent and accurate across the three coefficients and the parameters that influence the DEFF, save for some estimates that are slightly out of the range of simulation error. This is particularly important for the cross-level interaction term, β_3. Cross-level interactions are common in state politics research (e.g., Wolfinger, Highton, and Mullin 2005) and are unique because they exhibit both individual- and cluster-level variation. The results presented here as well as past work show that RCSE do not handle this characteristic well (Erikson, Pinto, and Rader 2010). In contrast, BCSE coverage probabilities show virtually no bias for these variables.

Applications of Clustered Data in Political Science

Having shown that BCSE perform well in a controlled setting, I turn next to evaluating their performance with real data. I replicate three analyses with BCSE: a randomized experiment on the impact of *Rock the Vote*'s television advertisements on turnout in the 2004 presidential election (Green and Vavreck 2008), an article on state-level factors affecting voter registration rates (Brown, Jackson, and Wright 1999), and a study on the role of incumbent voting behavior and electoral success in state legislatures (Hogan 2008).[19] The central objectives with these applications are to assess the external validity of the simulation study and to show that BCSE can make a substantive impact on state politics research.

Green and Vavreck (2008): Rock the Vote *and Voter Turnout*

In their recent analysis of the effect of clustered data on standard errors, Green and Vavreck show results from a randomized experiment investigating the effectiveness of get out the vote (GOTV) ads produced by *Rock the Vote* during the 2004 presidential election.[20] The authors matched 85 cable systems into 40 strata based on past turnout rate and assigned at least one system from each stratum to the treatment group.[21] Two different GOTV ads aired on various networks in the treatment cable systems, while the control systems did not air these ads. The data include 23,869 individuals 18 to 19 years old living in one of the 85 systems, and the dependent variable is a binary indicator of whether an individual in the study voted.

The authors regress this variable on an indicator for treatment and 39 indicators for all but one of the strata (i.e., stratum fixed effects).[22] However, because randomization occurred at the cable system-level rather than among individuals, there is reason to believe that there may be clustering in the data. Voters residing in the same system

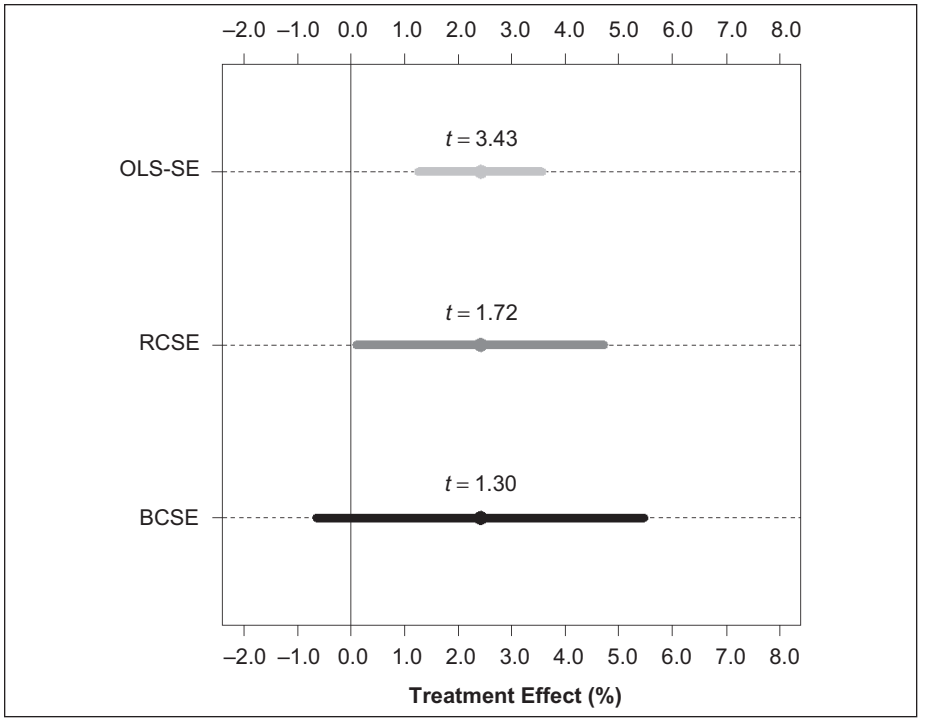

Figure 3. Reanalysis of the estimated treatment effect of *Rock the Vote* advertisements among 18- to 19-year-old voters (Green and Vavreck 2008, Table 3)
Note: The graph plots the coefficient estimate on the treatment variable and 90% confidence intervals calculated from each standard error method. Original results report OLS-SE and RCSE estimates.

experienced the same campaign environment, which may have produced any number of unmodeled factors affecting the decision to vote. To account for this possibility, the authors report both the OLS-SE and RCSE estimates, with the cable system as the cluster identifier. Figure 3 reports these results as well as the BCSE estimates in the form of the point estimate on the treatment effect and 90% confidence intervals. Note that one unique issue with these data stemming from the cluster-randomized design is the need to stratify the bootstrap samples. Specifically, in this case BCSE are calculated by resampling clusters *within each of the strata* to ensure that every stratum is represented in every bootstrap sample. This guarantees that the stratum fixed effects variables are not perfectly collinear with the intercept in a given bootstrap sample.

The OLS coefficient on the treatment variable (which varies only at the cluster level) is 2.4, indicating that "turnout was boosted by 2.4 percentage points among those living in cable systems assigned to the treatment group" (Green and Vavreck 2008, 150). However, the standard error estimates yield considerably different levels

of confidence in this result. The OLS-SE corresponds to a *t*-value of 3.43, which, as the authors note, is likely too large given the expectation of clustering in the data. The larger confidence interval constructed from RCSE provides some validation of this expectation, yielding a *t*-value of 1.72, which corresponds to statistical significance at the 90% confidence level. In contrast, if BCSE are used, the treatment effect can no longer be distinguished from zero at conventional levels of statistical significance, with a *t*-value of 1.30. However, it should also be noted that Green and Vavreck (2008) show additional models in which the treatment effect can be statistically distinguished from zero.[23]

Brown, Jackson, and Wright (1999): Democratic Party Control and State Voter Registration

Brown, Jackson, and Wright (1999) consider an intervening step between Democratic Party organization in the states and voter turnout: voter registration. Because partisan supporters can be effective only if they are registered to vote, the authors hypothesize that greater control by the Democratic Party in the state will correspond to an increase in voter registration. They reason that Democrats have more incentive than Republicans to make registration easier as a means of mobilizing the "have-nots" of the state, and thus states with greater Democratic Party control will evidence higher levels of registration.

The authors test this hypothesis by regressing state voter registration rates in the four federal election years from 1984 to 1990 on the Hill and Leighley (1996) measure of *liberal party control*. Their hypothesis predicts a positive coefficient on this variable. However, because there are repeated observations from each state in their model, it is reasonable to expect cluster correlation at the state level. As Arceneaux and Nickerson (2009, 184) note, "Voters in a state share the same political history, constellation of media markets, and set of statewide political elites—all of which are distinctive across states and difficult to model."

In addition, because the dependent variable is a continuous measure, this expectation can be validated through an empirical estimate of ρ. Several statistical software packages, including R and Stata, will estimate the intracluster correlation coefficient from the residuals of an OLS model (see the online appendix for sample code). The estimate from the model in column 3 of Table 1 in Brown, Jackson, and Wright (1999, 469) is 0.78. With 180 observations in 45 clusters, this corresponds to a DEFF of 3.34, indicating the presence of downward bias to standard errors that do not account for clustering. Figure 4 reports the model results, including point estimates and 95% confidence intervals constructed from the authors' original heteroscedasticity-robust HWSE, RCSE, and BCSE.[24]

The results provide another example of how using different standard error methods can lead to different interpretations. The coefficient on liberal party control is statistically significant at the 95% level if the HWSE ($t = 3.92$) or RCSE ($t = 2.23$) are used but is just on the edge of significance with BCSE ($t = 1.94$). The coefficient on the

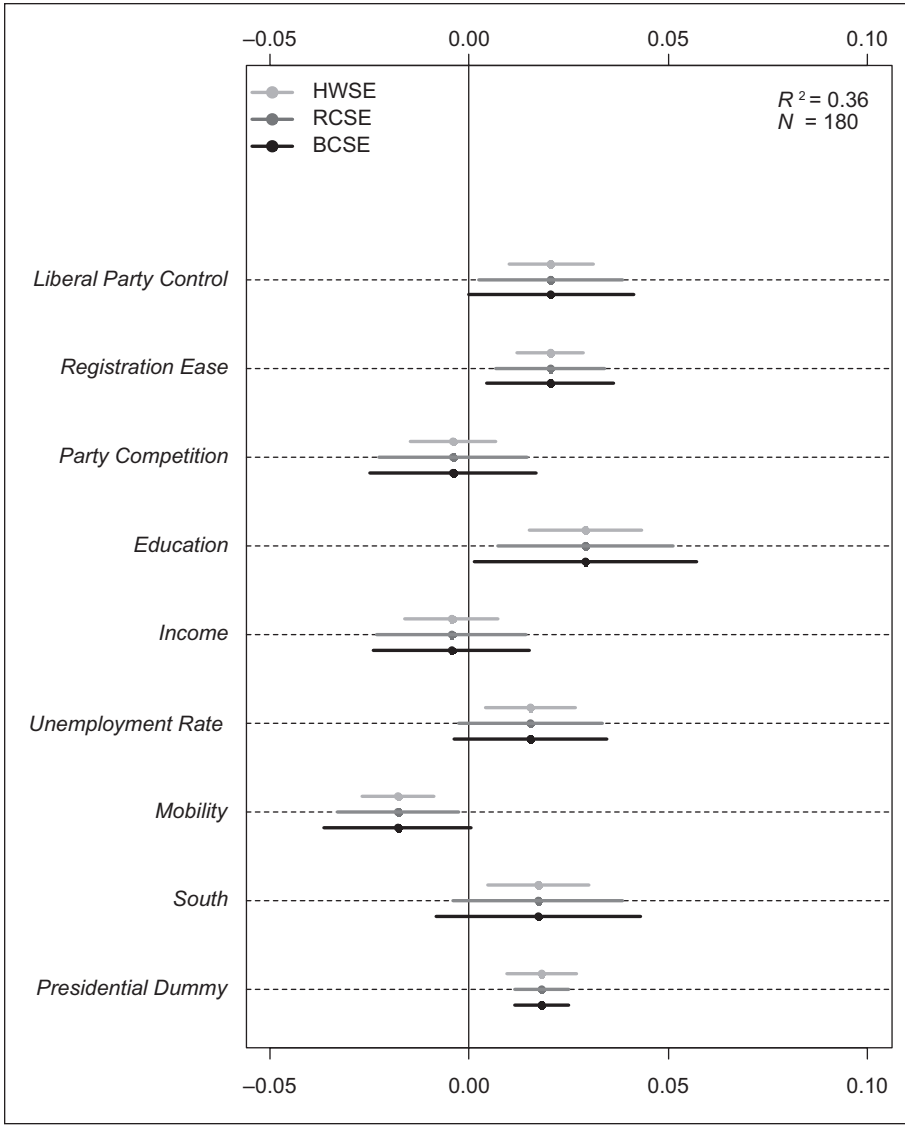

Figure 4. Reanalysis of factors affecting state voter registration, 1984–1990 (Brown, Jackson, and Wright 1999, Table 1)
Note: The graph plots standardized coefficient estimates and 95% confidence intervals calculated from each standard error method. Original results report HWSE estimates.

control for population mobility (*mobility*) shows a similar pattern; the *t*-values for the HWSE, RCSE, and BCSE estimates on that coefficient are 4.02, 2.30, and 1.91, respectively. Of course, significance levels are arbitrary, and in the case of liberal party

control the authors have good reason to interpret a *t*-value of 1.94 as support for their hypothesis. Nonetheless, this model shows that there can be differences between standard error methods that do and do not account for clustering *and* differences within the subset of methods that do account for clustering.

Hogan (2008): State Legislative Incumbent Responsiveness and Electoral Success

Hogan (2008) examines how the voting behavior of state legislators influences their chances of reelection. In particular, he expects that under certain conditions extreme partisan voting by incumbents will have either a negative or positive impact on the proportion of votes they receive. The data come from 1,816 incumbents in both the lower and upper houses of 14 states in 1996 and 1998. In the current analysis, I replicate the OLS model predicting incumbent two-party vote share (Hogan 2008, 867, Table 3).

The model contains three independent variables of interest along with several controls. *Partisan policy position* is a measure of incumbent divergence from expected district preferences on economic and regulatory policy. A larger value signifies a legislator whose voting record is strongly divergent from district preferences and more in line with the party. Hogan interacts this variable with measures of *challenger spending* and *incumbent spending*. He tests two hypotheses with this specification. First, as challenger spending increases, he hypothesizes that the effect of partisan policy position should decrease. In other words, he expects that as challengers spend more the electorate should become more informed about the incumbent's out-of-step voting record, and thus more partisan voting should correspond with a decline in incumbent vote share. However, he expects this to be counteracted by incumbent spending, hypothesizing that as that variable increases, the marginal effect of partisan policy position will also increase because the party base will be more informed and enthusiastic about the incumbent's partisan record (see Hogan 2008, 860–61).

As in the previous examples, a key issue with this study is the clustered nature of the data. If the incumbents in the sample are independent, OLS-SE will provide unbiased estimates of coefficient variability. 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, are subject to 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. The estimate of ρ for the model is 0.046, and with 1,816 observations grouped in 14 states, this corresponds to a DEFF of 10.22. Thus, conventional OLS-SE estimates, which Hogan uses for hypothesis testing, are likely to be biased downward.

Figure 5 reports the results with point estimates and 95% confidence intervals from his original OLS-SE estimates and the RCSE and BCSE estimates. In almost all cases BCSE are the largest of the three. The difference between the estimates is especially

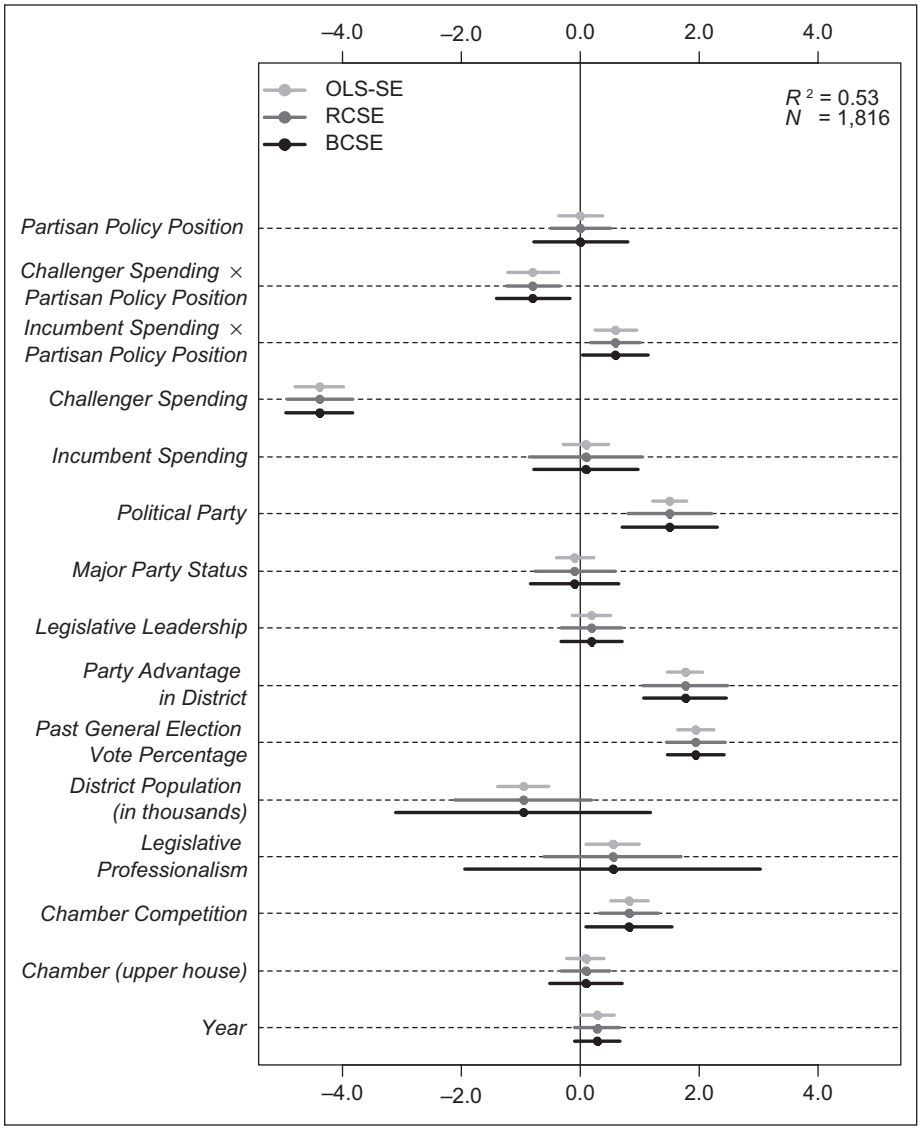

Figure 5. Reanalysis of factors affecting percentage of two-party vote received by challenged incumbents (Hogan 2008, Table 3)
Note: The graph plots standardized coefficient estimates and 95% confidence intervals calculated from each standard error method. Original results report OLS-SE estimates.

large for the coefficient on *legislative professionalism*. That variable varies only at the cluster level (states), and thus it is not surprising that the BCSE estimate is much larger than the other two in light of the simulation results.

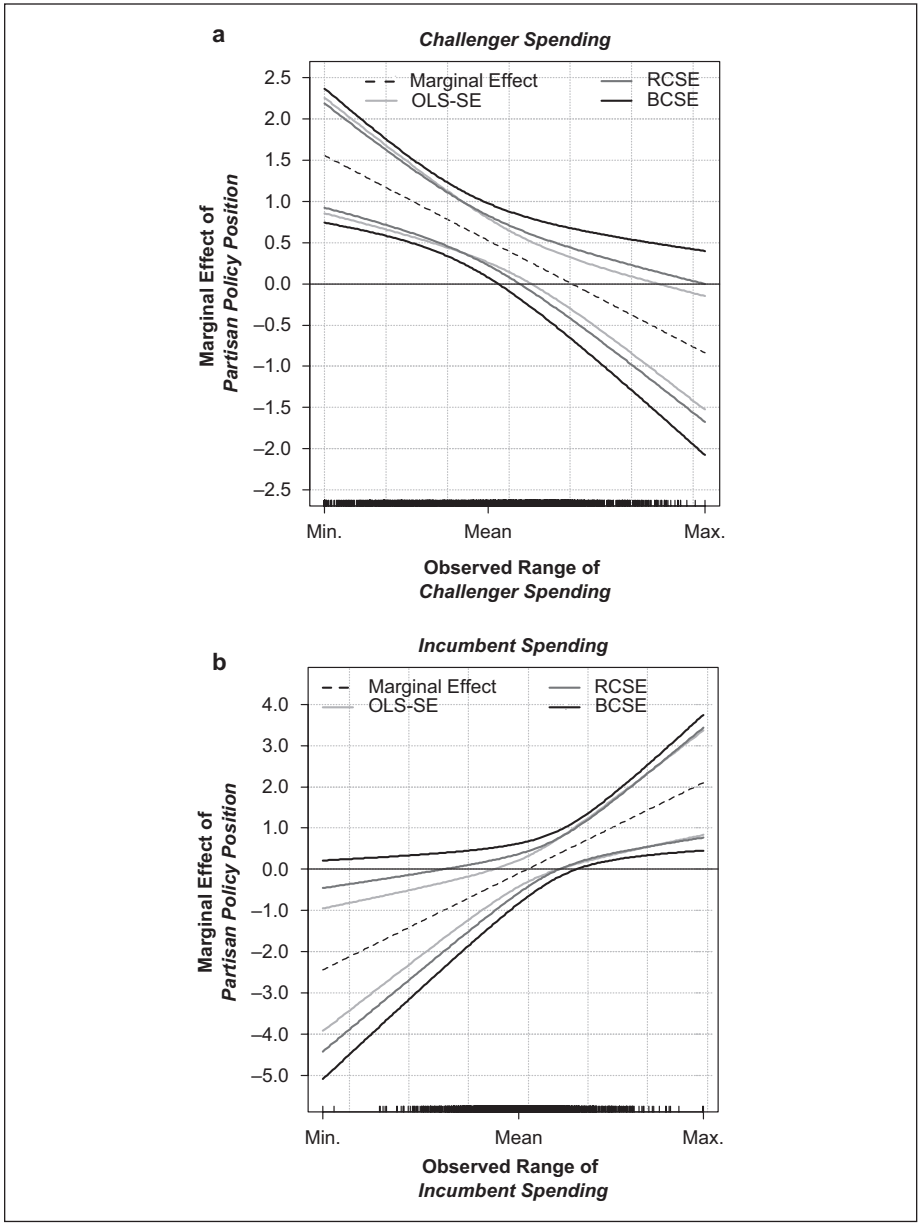

Figure 6. The marginal effect of partisan policy position, conditioned by challenger spending and incumbent spending (Hogan 2008)

Note: The graphs plot the marginal effect of partisan policy position (black dashed lines) across the observed ranges of challenger spending (panel a) and incumbent spending (panel b) and 95% confidence intervals calculated from the three standard error methods (solid lines). Notice that the marginal effect is never negative and statistically significant if the BCSE method is used for hypothesis testing.

However, because the hypotheses posit interactive effects, a standard report of the coefficients as they are displayed in Figure 5 is insufficient for understanding the effects of those variables (Brambor, Clark, and Golder 2006). Figure 6 displays the marginal effect of partisan policy position on the y-axes across the range of challenger spending (panel a) and incumbent spending (panel b) on the x-axes along with 95% confidence intervals constructed from each of the three standard error methods examined here. Those graphs show more clearly the substantive implications of using BCSE.

Consistent with expectations, the marginal effect of partisan policy position decreases as challenger spending increases and increases along with incumbent spending, but the three confidence interval estimates provide different pictures of uncertainty. The OLS-SE and RCSE estimates indicate that the effect is negative and statistically significant at the 95% level for some segment of the x-axes. This is supportive of the hypotheses outlined above: given certain conditions (high spending by challengers and/or low spending by incumbents) a partisan voting record can have a negative impact on an incumbent's vote share. However, with the BCSE estimates, the marginal effect of partisan policy position is never negative and statistically significant. Thus, accounting for clustering with BCSE leads to somewhat less support for the original hypotheses than with OLS-SE or RCSE.

Assessing the Replication Results

Many of the patterns evident in the simulations appear again in the replications. For instance, BCSE are the largest standard error estimates in the simulations and almost always the largest estimates in the replications. Of course, this would be a problem if BCSE were biased upward, but the simulations show that this is generally not the case—BCSE are the largest because they are unbiased while OLS-SE and RCSE are too small. Thus, the fact that BCSE are the largest on average in the replications indicates those estimates should be trusted more.

In addition, the simulation results show that there is a small amount of difference between the three standard error methods for individual-level variables and bigger differences for variables that exhibit cluster-level variation (see Figure 1). That finding appears again in the replications. Consider the partisan policy position (individual level) and legislative professionalism (cluster level) variables in the Hogan (2008) model. The ratio between the BCSE and OLS-SE estimates for the latter is much larger (5.58) than that of the former (2.15). Overall, the replications provide external validation of the simulation study and show that the difference between BCSE and the other standard error methods can have implications for substantive conclusions in state politics research.

Conclusions

The results reported here indicate that researchers should be concerned with the presence of clustering because its effects can hinder their ability to conduct proper

statistical inference. However, that does not mean researchers should avoid clustered data. In such a situation, I recommend the use of BCSE rather than OLS-SE or RCSE. While RCSE are common in political science, simulations detailed here and elsewhere indicate that they are often too small. In contrast, the results show convincingly that BCSE provide a better estimate of coefficient variability even in the presence of a large DEFF. The method is easy to implement and hypothesis testing procedures do not change. Thus, researchers should be very confident in their ability to conduct proper statistical inference when using BCSE.

In addition, note the common pattern across all of the simulation results that adding clusters reduces the DEFF, which reduces standard error bias. Researchers who have control over the data collection process would be better off adding new 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 DEFF, but adding more observations within existing clusters simply increases the magnitude of the DEFF. For example, a state politics researcher could optimize a data collection process by gathering fewer observations from several different states rather than a large number from only a few states. However, researchers often do not have any control over the data collection process and might be forced to use data with many observations and few clusters. The results described here indicate BCSE will still allow the analyst in that situation to obtain accurate estimates of coefficient variability.

Appendix

As is mentioned in the main text, there are several reasons that the BCSE method may be preferable to the CLRT method advocated by Erikson, Pinto, and Rader (2010). More specifically, CLRT does not calculate a full covariance matrix of the parameter estimates, is driven by a much different philosophy than the other methods I examine, and is more difficult to implement than BCSE. I address the first two of these reasons in more detail here. See the online appendix for additional details.

Covariance Matrix versus Hypothesis Testing

Randomization testing was designed specifically to get a p value for a statistic of interest (Fisher 1922). This is different from estimation of a covariance matrix (and thus, standard errors), which is a complete representation of model uncertainty that can be used to get a p value. Erikson, Pinto, and Rader (2010) obtain standard errors out of their CLRT procedure by calculating the variance of the resampled estimates, but the method does not calculate covariance between coefficients. This is problematic because a covariance estimate is often needed for analysis of model results. For instance, conducting joint F tests, calculating confidence intervals for the marginal effects of variables in interaction models (e.g., Figure 6), and the commonly used CLARIFY procedure of King, Tomz, and Wittenberg (2000) all require an estimate of

covariance between coefficients of interest. The BCSE method provides an estimate of the full covariance matrix.

Philosophy

CLRT also requires researchers to change how they view error in their models. The general method of randomization testing conceives of the stochastic element of a model as error fixed (in repeated samples) to observations. This differs from the traditional econometric view that the stochastic element is additive error that would not necessarily be the same for a given subject if the data-generating process were repeated. In other words, in the traditional view that most political scientists learn, error enters "via repeated drawing of individual error terms," while CLRT conceives of error as entering the model from "repeated random assignments of treatments to subjects" (Kennedy 1995, 86). The CLRT view is not wrong, but it is different, which might not be appealing to applied researchers.[25]

Acknowledgments

I thank Tom Carsey for excellent guidance over the course of this project. I also appreciate comments and assistance from Fred Boehmke, Skyler Cranmer, Bruce Desmarais, Donald Green, Matt Golder, Shubin Liu, Jim Stimson, Lynn Vavreck, Nick Weller, and members of Triangle Political Methodology Group. Finally, I thank Donald Green and Lynn Vavreck, Robert Jackson, and Robert Hogan for making replication data and code available. A previous version of this article was presented at the Ninth Annual Conference on State Politics and Policy, May 22, 2009, Chapel Hill, NC. Simulation code, full results, and an online appendix are available at http://academic.udayton.edu/SPPQ-TPR/index.htm. Replication code of the substantive examples will be made available with permission from the original author(s). All errors are my own.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Time-series cross-section (TSCS) data could also be categorized as clustered data but are not considered in this study because they can have issues that cross-sectionally clustered data do not have, such as autocorrelation. See the special issue of *Political Analysis* (vol. 15, no. 2) for leading work on TSCS data.
2. Other names for the method include "cluster bootstrap," "case bootstrap," "nonparametric bootstrap," and "nonoverlapping block bootstrap" (Cameron, Gelbach, and Miller 2008, 416).

3. In the case of unbalanced data, the design effect (DEFF) can also be calculated as $1+ \rho\cdot(\sum_{c=1}^{C} n_c^2 /N - 1)$, where n_c is the number of observations in cluster c . This reduces to equation 2 when the clusters are balanced.
4. The DEFF is commonly calculated in experimental research when there exists only one covariate of interest—the treatment effect. In the case of a regression with multiple independent variables, it can affect each variable differently, but this variability is generally small, and thus the DEFF is still informative (Alecxih and Corea 1998, 5).
5. In fact, in some cases RCSE can be the same size or smaller than OLS-SE. As Arceneaux and Nickerson (2009) note, it is possible for unexplained variance to be greater within a cluster than between clusters (i.e., negative cluster correlation). The RCSE method incorporates the level of variation within clusters into its estimate, and if variance is larger between clusters than it is within, the RCSE estimate will be smaller than that of OLS-SE.
6. Furthermore, the BCSE method can be extended to both of these estimators (Harden 2010).
7. In addition, the BCSE method is easier to implement in statistical software. See note 9 and the appendix for more details on these points.
8. In fact, if each observation is its own cluster (i.e., $C = N$ and $n_c = 1$), then RCSE become HWSE (Primo, Jacobsmeier, and Milyo 2007, 451).
9. More specifically, CLRT assumes that, conditional on the covariates in the model, observations in one cluster are drawn from the same distribution as observations in another cluster. This is a stronger supposition than simply assuming observations in different clusters are independent, as with BCSE.
10. I conducted the simulations in R version 2.8.1 (R Development Core Team 2008) with the Design (Harrell 2008a), Hmisc (Harrell 2008b), mvtnorm (Genz et al., 2008), and sandwich (Zeileis 2006) packages and a function for RCSE created by Arai (2009). I then confirmed the results in Stata 10/SE (StataCorp 2007). BCSE were estimated with $B = 1{,}000$.
11. The simulation design creates an equal number of observations in each cluster. Switching to an unbalanced design does not change conclusions.
12. I also simulated data with $\rho = .00$ (i.e., no clustering). All three methods performed well in this case, though RCSE were slightly too small.
13. Results are unchanged if another value, such as 0.50, is set as the standard.
14. I use this method because it is common in the statistics literature and because of its simple interpretation, but results are not dependent on the choice. For example, evaluating standard error performance by comparing the mean standard error of each coefficient to the standard deviation of the simulated estimates of each coefficient does not change the substantive conclusions.
15. I also repeated the process with six different random number generators; results were virtually identical in each case.
16. Full results are available at http://academic.udayton.edu/SPPQ-TPR/index.htm.
17. The true 95% simulation error bounds for a given coverage probability estimate, $\hat{p}$**,** would be $\hat{p} - 2\cdot\sqrt{\hat{p}\cdot(1-\hat{p})/10000}$. Plotting lines only at $0.95 \pm 2\ .\ \sqrt{0.95\cdot 0.05/10000}$ as I do allows for more visual clarity.
18. Specifically, I conducted the simulations with an "artificial" version of RCSE in which I used the true model error (ε) instead of the residuals ($\hat{\varepsilon}$) in the calculation. In other words, I

calculated the covariance matrix exactly as in equation 4, but inserted ε_i instead of $\hat{\varepsilon}_i$ into the "meat" matrix. This manipulation drastically improves the RCSE estimates, but of course is never available in an applied setting. See the full results for details.

19. I replicated each model exactly in R version 2.8.1 (R Development Core Team 2008) with the Design, (Harrell 2008a) Hmisc (Harrell 2008b), sandwich (Zeileis 2006), and lmtest (Zeileis and Hothorn 2002) packages and the Arai RCSE function (Arai 2009). I then confirmed the results in Stata 10/SE (StataCorp 2007). The BCSE in each model were estimated with $B = 10{,}000$.
20. For a reanalysis of this study that comes to a similar conclusion using Bayesian hierarchical modeling, see Jackman (2009, 355).
21. Five strata contained three systems each and the other thirty-five contained two each. Three of the three-system strata were composed of one control and two treatment systems and the other two contained two control and one treatment system. The other thirty-five strata contained one treatment and one control in each.
22. Estimation with logistic regression or another binary dependent variable model does not change results. Thus, I maintain the original authors' choice of OLS.
23. Despite the fact that BCSE lead to a null finding with this particular model, it is important to note that the authors still find strong evidence that the treatment had a positive effect on turnout. They report results from aggregate-level OLS, in which the clustering is entirely removed from the data, which produces the same coefficient estimate of approximately 2.4% and a *t*-value of 1.57. Furthermore, the generalized least squares coefficient estimate on the treatment variable is about 3.0% with a *t*-value of 2.12 (see Green and Vavreck 2008, 149). Finally, randomization testing, which is the most appropriate method of conducting inference in a randomized experiment, produces a *p* value of .098 on the treatment effect of 2.4%.
24. HWSE performance in clustered data is virtually identical to that of OLS-SE (Green and Vavreck 2008, 150).
25. In addition, recall that CLRT assumes independence and exchangeability of errors between clusters while BCSE assumes only independence. Researchers never know if the exchangeability assumption holds, so a method that does not require it is preferable to one that does.

References

Alecxih, Lisa, and John Corea. 1998. "Deriving State-Level Estimates from Three National Surveys: A Statistical Assessment and State Tabulations." http://aspe.hhs.gov/daltcp/reports/deriving.pdf (Accessed May 15, 2009).

Arai, Mahmood. 2009. "Cluster-Robust Standard Errors Using R." http://people.su.se/~ma/clustering.pdf (Accessed May 29, 2009).

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.

Arceneaux, Kevin, and Gregory Huber. 2007. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science* 51: 957–77.

Arceneaux, Kevin, and David W. Nickerson. 2009. "Modeling Certainty with Clustered Data: A Comparison of Methods." *Political Analysis* 17: 177–90.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14: 63–82.

Brown, Robert D., Robert A. Jackson, and Gerald C. Wright. 1999. "Registration, Turnout, and State Party Systems." *Political Research Quarterly* 52: 463–79.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90: 414–27.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

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

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

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Erikson, Robert S., Pablo M. Pinto, and Kelly T. Rader. 2010. "Randomization Tests and Multi-Level Data in State Politics." *State Politics & Policy Quarterly* 10: 180–98.

Feng, Ziding, Dale McLerran, and James Grizzle. 1996. "A Comparison of Statistical Methods for Clustered Data Analysis with Gaussian Error." *Statistics in Medicine* 15: 1793–1806.

Fisher, Ronald A. 1922. "On the Interpretation of x^2 from Contingency Tables, and the Calculation of p." *Journal of the Royal Statistical Society* 85: 87–94.

Franzese, Robert J. 2005. "Empirical Strategies for Various Manifestations of Multilevel Data." *Political Analysis* 13: 430–46.

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

Genz, Alan, Frank Bretz, Torsten Hothorn, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, and Fabian Scheipl. 2008. "mvtnorm: Multivariate Normal and *t* Distributions." R package version 0.9-3.http://CRAN.R-project.org/package=mvtnorm (Accessed May 29, 2009).

Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Techniques." *Political Analysis* 16: 138–52.

Harden, Jeffrey J. 2010. "Improving Statistical Inference with Clustered Data." University of North Carolina at Chapel Hill. Typescript.

Harrell, Frank E. 2008a. "Design: Design Package." R package version 2.1-2. http://biostat.mc.vanderbilt.edu/s/Design (Accessed May 29, 2009).

Harrell, Frank E. 2008b. "Hmisc: Harrell Miscellaneous." R package version 3.4-4. http://biostat.mc.vanderbilt.edu/s/Hmisc (Accessed May 29, 2009).

Hill, Kim Quaile, and Jan E. Leighley. 1996. "Political Parties and Class Mobilization in Contemporary United States Elections." *American Journal of Political Science* 40: 787–804.

Hogan, Robert E. 2008. "Policy Responsiveness and Incumbent Reelection in State Legislatures." *American Journal of Political Science* 52: 858–73.

Huber, Peter 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: University of California Press, 221–33.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. New York: John Wiley.

Kennedy, Peter E. 1995. "Randomization Tests in Econometrics." *Journal of Business & Economic Statistics* 13: 85–94.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44: 341–55.

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

Künsch, Hans R. 1989. "The Jackknife and the Bootstrap for General Stationary Observations." *Annals of Statistics* 17: 1217–41.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73: 13–22.

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: 334–38.

Primo, David M., Matthew L. Jacobsmeier, and Jeffrey Milyo. 2007. "Estimating the Impact of State Policies and Institutions with Mixed-Level Data." *State Politics & Policy Quarterly* 7: 446–59.

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

R Development Core Team. 2008. "R: A Language and Environment for Statistical Computing." Vienna, Austria:R Foundation for Statistical Computing. http://www.r-project.org.

Rogers, William H. 1993. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 13: 19–23.

StataCorp. 2007. "Stata Statistical Software: Release 10." College Station, TX: StataCorp.

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 & Policy Quarterly* 3: 23–41.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48: 817–38.

Williams, Rick L. 2000. "A Note on Robust Variance Estimation for Cluster-Correlated Data." *Biometrics* 56: 645–46.

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

Zeileis, Achim. 2006. "Object-Oriented Computation of Sandwich Estimators." *Journal of Statistical Software* 16: 1–16.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." R News 2: 7–10. http://CRAN.R-project.org/doc/Rnews/May 29, 2009). 2: 7–10.

Zorn, Christopher. 2001. "Generalized Estimating Equation Models for Correlated Data: A Review with Applications." *American Journal of Political Science* 45: 470–90.

Zorn, Christopher. 2006. "Comparing GEE and Robust Standard Errors for Conditionally Dependent Data." *Political Research Quarterly* 59: 329–41.

Bio

Jeffrey J. Harden is a PhD candidate in political science at the University of North Carolina at Chapel Hill. His research focuses primarily on U.S. state politics with a particular emphasis in public opinion, representation, and methodological issues.