PA

Corrected Standard Errors with Clustered Data

John E. Jackson

Department of Political Science, University of Michigan, Ann Arbor, MI 48109, USA. Email: jjacksn@umich.edu

Abstract

The use of cluster robust standard errors (CRSE) is common as data are often collected from units, such as cities, states or countries, with multiple observations per unit. There is considerable discussion of how best to estimate standard errors and confidence intervals when using CRSE (Harden 2011; Imbens and Kolesár 2016; MacKinnon and Webb 2017; Esarey and Menger 2019). Extensive simulations in this literature and here show that CRSE seriously underestimate coefficient standard errors and their associated confidence intervals, particularly with a small number of clusters and when there is little within cluster variation in the explanatory variables. These same simulations show that a method developed here provides more reliable estimates of coefficient standard errors. They underestimate confidence intervals for tests of individual and sets of coefficients in extreme conditions, but by far less than do CRSE. Simulations also show that this method produces more accurate standard error and confidence interval estimates than bootstrapping, which is often recommended as an alternative to CRSE.

Keywords: clustered robust standard errors, clustered data, confidence intervals

Common units of analysis in Political Science are collective entities, such as voting precincts, cities, states or countries because that is how available data are collected, as in ecological voting studies, or because the unit is the entity of interest, as in studies of electoral party systems. Data clustered this way present estimation problems because the error terms within a cluster are unlikely to be independent. In many studies the multiple observations per unit are observed at irregular intervals with varying numbers of observations per unit so the data do not conform to the standard time-series, cross-section estimation models, which is one method for accommodating interdependence. The interdependence problem, however, must be addressed to obtain reliable estimates of the coefficients' uncertainty as measured by their standard errors.

The common method with clustered data is cluster robust standard errors (CRSE), based on the Liang and Zeger (1986) extension to clustered data of the robust standard errors associated with Eicker (1967), Huber (1967) and White (1980). Unfortunately, as reviewed in various studies cited in Imbens and Kolesár (2016), MacKinnon and Webb (2017) and Esarey and Menger (2019) CRSE are biased downwards for small samples and possibly even for larger samples. These authors present methods to adjust the CRSE confidence intervals, but do not address the basic bias problem.

This paper contributes to this literature by discussing a method for using the residuals from an OLS estimation to estimate the covariance within clusters which is then used to estimate the coefficient variance–covariance matrix. The problems with CRSE are discussed and then the new approach is developed. Extensive Monte Carlo simulations are conducted to compare the different correction methods with different sample properties. We conclude with examples from the State Politics and the Comparative Politics literatures where clustered data are common.[1]

1 Methodological Issues with Cluster Robust Standard Errors

CRSE are routine with grouped data. (See Greene (2012, pp. 351 and 353); Franzese (2005); and the citation count in Esarey and Menger (2019, Table 1).) MacKinnon and Webb (2017, p. 233)

Author's note: I want to thank Ken Kollman, Chuck Shipan, Matthew Webb and the ubiquitous anonymous referee for their helpful comments and Diogo Ferrari for his comments and the R package "ceser" for computing CESE. All are absolved from any and all errors.

1 Replication files and data for all simulations and examples are archived at Jackson (2019).

Political Analysis (2020)
vol. 28:318–339
DOI: 10.1017/pan.2019.38

Published
15 October 2019

Corresponding author
John E. Jackson

Edited by
Jeff Gill

name three necessary conditions for CRSE to be consistent: (a) an infinite number of clusters; (b) homogeneity across clusters in the stochastic term distributions; and (c) an equal number of observations per cluster. This section discusses these and additional conditions that affect the CRSE underestimates of the true standard errors and thus biases subsequent statistical inference.

1.1 OLS with Cluster Robust Standard Errors

The problems posed by clustered data are described well in the cited literature. Following the conventional notation the population model is $Y = X\beta + V$ and the estimated model is $Y = Xb + e$. The observations used to obtain the estimated model are clustered into groups, such as countries. The clusters are denoted by $g, g = 1, \ldots, G$. With interdependence within clusters $E(V_g V_g') = \Sigma_g \neq \sigma_g^2 I$. Denote the off-diagonal terms in Σ_g by $\rho_g \neq 0$. The conventional assumption and the one adopted here is that the stochastic terms are independent across groups so that $\Sigma_v = E(VV')$ is block diagonal with the blocks being the variances and covariances of the stochastic terms within each cluster. The correct standard errors for the OLS estimation are,

$$\Sigma_b = (X'X)^{-1}\left[\sum_{g=1}^{G}(X_g'\Sigma_g X_g)\right](X'X)^{-1}, \tag{1}$$

and are estimated with the expression,

$$\hat{\Sigma}_b = S_b = (X'X)^{-1}\left[\sum_{g=1}^{G}(X_g'\hat{\Sigma}_g X_g)\right](X'X)^{-1}, \tag{2}$$

where $\hat{\Sigma}_g$ is an estimate for Σ_g.

Cluster robust standard errors, following the procedure for sandwich estimators, obtain $\hat{\Sigma}_g$ from the OLS residuals, $\hat{\Sigma}_g = (e_g e_g')$, giving the following expression for the estimated CRSE,

$$S_b = (X'X)^{-1}\left[\sum_{g=1}^{G} X_g'(e_g e_g')X_g\right](X'X)^{-1}. \tag{3}$$

Comparison of the true standard errors to the expected CRSE is a comparison of the bracketed term in Equation (1) with the expected value of the bracketed term in Equation (3), which in turn depends on $E(e_g e_g')$. Appendix A develops the following expression,

$$E(e_g e_g') = \Sigma_g - \Sigma_g P_g - P_g\Sigma_g + X_g(X'X)^{-1}\left(\sum_{g=1}^{G} X_g'\Sigma_g X_g\right)(X'X)^{-1}X_g', \tag{4}$$

where $P_g = X_g(X'X)^{-1}X_g'$. Equation (4) shows the expected value of the variance–covariance matrix used to estimate Σ_g. Unfortunately substituting Equation (4) into the expected value of Equation (3) does not lead to any easily interpretable expression. Appendix B does this for the bivariate case assuming homogeneous clusters, i.e. $\Sigma_g = \Sigma$ for all g. The results provide some important insights into factors that affect by how much CRSE underestimate Σ_b.

One key result in Appendix B is Equation (B 5), which shows for the bivariate case that,

$$\sigma_b^2 = (x'x)^{-1}\left(\sum_g x_g\Sigma_g x_g\right)(x'x)^{-1} = \frac{\sigma^2}{NV_x} + \frac{2\rho}{(NV_x)^2}\sum_g C_g^*, \tag{5}$$

where x denotes the deviations of the explanatory variable from its full sample mean, $x = (X - \bar{X})$; N is the total sample size; V_x is the full sample variance of the explanatory variable X, and C_g^* is the sum of the cross-product terms in cluster g, $C_g^* = \sum_{i=1}^{n_g-1} \sum_{j=i+1}^{n_g} x_{gi} x_{gj}$. Equation (B 4) in Appendix B shows that the expression for the expected value of the CRSE estimated coefficient variance is

$$E(s_b^2) = \frac{\sigma^2}{NV_x}\left[1 - \sum_g \left(\frac{n_g V_{x_g}}{NV_x}\right)^2\right] + \frac{2\rho}{(NV_x)^2}\sum_g \left[\left(1 - \frac{n_g V_{x_g}}{NV_x}\right)^2 C_g^*\right]. \tag{6}$$

There are several important implications that can be drawn from Equations (5) and (6). The difference between the two equations is the presence of the squared terms involving $n_g V_{x_g}/NV_x$, which is the share of the total variance in X contributed by cluster g. The larger these summation terms the larger the amount by which the CRSE underestimate the true standard errors. Though Equation (6) is derived from the bivariate model and does not translate directly to the general multivariate model it still provides insight into factors that likely affect the reliability of CRSE in the general case.

There are several factors that affect the magnitude of the sum of these squared terms. One is the homogeneity of the distribution of X across clusters. Appendix B.1 shows that with homogeneous clusters, meaning that the number of observations and the mean and variance of X are identical for all clusters, then the terms in brackets are only related to the factor $1/G$ so that as the number of clusters increases the CRSE approach the true standard errors, $\lim_{G\to\infty} s_b = \sigma_b$.

Cluster heterogeneity is the more frequent and interesting case. Because the relevant terms are squared quantities increasing heterogeneity increases the sum of these terms, which increases the gap between the CRSE and the true standard errors. One obvious source of heterogeneity is variation in the number of observations in each cluster, n_g. As this variation increases so will the bias in the CRSE. Equal cluster sizes are one of the necessary conditions for consistent CRSE.

Heterogeneity in the explanatory variables' variance is a much less discussed problem, though this heterogeneity has many sources. The focus here is on the amount of between cluster variance as a proportion of the total variance in an explanatory variable, labeled the amount of covariate clustering.[2] A high degree of covariate clustering is a likely attribute of many Political Science clustered datasets where the majority of variation in entities of interest, such as electoral rules, institutions, etc. is between units, such as countries, rather than within units. To see the effect of increasing the between variance decompose $n_g V_{x_g}$ into its within and between components,

$$n_g V_{x_g} = \sum_{i=1}^{n_g}[(X_{gi} - \bar{X}_g)^2 + n_g(\bar{X}_g - \bar{X})^2 = n_g v_{x_g} + n_g(\bar{X}_g - \bar{X})^2], \tag{7}$$

where v_{x_g} is the variance of X within cluster g. Increased covariate clustering means larger values for the $n_g(\bar{X}_g - \bar{X})^2$ component, which in turn increases the variation in the $n_g V_{x_g}$ terms in Equation (6), which in turn decreases the value for $E(s_b^2)$ relative to σ_b^2. As the amount of covariate clustering increases the gap between the CRSE and the true standard error also increases.

2 Corrections

What are the alternatives? Bell and McCaffrey (2002), based on a recommendation in Davidson and MacKinnon (1993), suggest an adjustment to the residual variances that increases the coefficient

2 Harden (2011, footnote 3) discusses covariate clustering but combines it with a term assessing the variation in observations per cluster, conflating two sources of heterogeneity. His exact expression is, $1+\rho[(1/N)(\sum_{g=1}^{G} n_g^2)-1]$, where ρ is the amount of covariate clustering.

standard errors.[3] Imbens and Kolesár (2016) incorporate this adjustment in their method. This approach still relies on $e_g e_g'$ to estimate Σ_g. Following the sandwich metaphor used to describe Equation (3) this method alters the bread but not the meat, which is where the problem lies.[4]

Two methods adjust the confidence intervals for each coefficient, but not the standard errors, hoping to improve statistical inferences on a coefficient by coefficient basis. Imbens and Kolesár (2016) and Stata adjust the number of degrees of freedom for each coefficient, thereby expanding the confidence interval. Another approach uses bootstrap methods to estimate the confidence intervals, (Cameron, Gelbach, and Miller 2008; Harden 2011; MacKinnon and Webb 2017). Esarey and Menger (2019) have an excellent summary and comparison of different bootstrap methods.

These approaches are limited because they do not provide an estimate for Σ_b.[5] Exceptions are Harden (2011) who uses a bootstrapping procedure that returns an estimate for Σ_b and the Stata boottest program, Roodman *et al.* (2019).[6] We return to Harden in a later section. By far the most common correction though follows Cameron, Gelbach, and Miller (2008) and only corrects the individual coefficient confidence intervals, which creates serious limitations. A major one is that tests of null hypotheses about multiple coefficients, such as in models with interaction terms, are infeasible. This is a serious shortcoming given the popularity of models with interaction terms and the associated marginal effects plots, which require the full Σ_b matrix to compute the desired confidence intervals. The corrections focused on confidence intervals also constrain the discussion to the confidence intervals selected by the authors as computing intervals for different α values requires the original data. This is contrary to much current practice of reporting coefficients and standard errors and letting readers choose the level of uncertainty, see Wasserstein, Schirm, and Lazar (2019). A better strategy is to estimate the correct standard errors, which is done next.

2.1 Cluster Estimated Standard Errors

The proposed method begins with the expression for $E(e_g e_g')$. The strategy is to relate this expected value to the unobserved variances, $\sigma^2_{g_i}$, and covariances among pairs of stochastic terms, $\rho_{g_{ij}}$, and then to use these expressions to derive estimates for $\hat{\Sigma}_g$ in computing $\hat{\Sigma}_b$ in Equation (2).

Begin with the expression for $E(e_g e_g')$ in Equation (4). Impose the condition that the stochastic terms within a cluster are identically distributed so that $\sigma^2_{v_{g_i}} = \sigma^2_g$ and that $\rho_{g_{ij}} = \rho_g$ for all i and j in g. Equation (A 2) in Appendix A shows that,

$$\begin{aligned} E(e_g e_g') &= \sigma^2_g (I_g - P_g) + \rho_g \left[\iota_g \iota_g' - (I_g - P_g) - (P_g \iota_g \iota_g' + \iota_g \iota_g' P_g) \right. \\ &\quad \left. + X_g (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \iota_g \iota_g' X_g \right) (X'X)^{-1} X_g' \right] \\ &= \sigma^2_g Q_{1_g} + \rho_g Q_{2_g}, \end{aligned} \tag{8}$$

where ι_g is a $n_g \ x \ 1$ column vector of ones. Equation (8) shows that the expected value of the squares and cross-products of the residuals in each cluster are linear functions of the unknown

3 Davidson and MacKinnon refer to this as the hc_2 adjustment. See Section 2.1.

4 A few early Monte Carlo simulations showed that the Bell and McCaffrey method provides only slight improvements over CRSE so to conserve space it was not pursued.

5 Esarey and Menger (2019) examine a method developed by Ibragimov and Muller (2002) called cluster-adjusted t-statistics that does provide an estimate for Σ_b. This method, however, requires the model be estimated separately for each cluster. This is impossible when the number of observations within a cluster is less than the number of explanatory variables or if there are variables that do not vary within a cluster.

6 Boottest returns the full coefficient variance–covariance matrix if the test command includes all the right hand side variables including the constant term.

terms σ_g^2 and ρ_g with the weighting terms being functions of the observed Xs, e.g. the elements in Q_{1_g} and Q_{2_g}.

To simplify the algebra while we illustrate the methodology we impose the homogeneity condition across groups, so that $\sigma_g^2 = \sigma^2$ and $\rho_g = \rho$ for all g. Estimates for σ^2 and ρ follow from the linear structure in Equation (8). There are $n_g(n_g + 1)/2$ residual squares and cross-products in $(e_g e_g')$ whose expected values are linear functions of σ^2 and ρ. Pooling these elements for all groups gives $\sum_{g=1}^{G} n_g(n_g + 1)/2$ elements in the matrix of observed residual squares and cross-products,

$$S_e = \begin{bmatrix} e_1 e_1' & 0 & \cdots & 0 \\ 0 & e_2 e_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_G e_G' \end{bmatrix}.$$

These elements are related to the two unknown coefficients we are estimating and the corresponding elements in Q_1 and Q_2. Think of this as a simple linear regression model,

$$\begin{aligned} S_e &= \begin{bmatrix} e_1 e_1' & 0 & \cdots & 0 \\ 0 & e_2 e_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_G e_G' \end{bmatrix} = \begin{bmatrix} E(e_1 e_1') + \xi_1 & 0 & \cdots & 0 \\ 0 & E(e_2 e_2') + \xi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E(e_G e_G') + \xi_G \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} Q_{1_1} & 0 & \cdots & 0 \\ 0 & Q_{1_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{1_G} \end{bmatrix} + \rho \begin{bmatrix} Q_{2_1} & 0 & \cdots & 0 \\ 0 & Q_{2_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{2_G} \end{bmatrix} + \begin{bmatrix} \xi_1 & 0 & \cdots & 0 \\ 0 & \xi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi_G \end{bmatrix} \\ &= \sigma^2 Q_1 + \rho Q_2 + \xi. \end{aligned} \tag{9}$$

Estimation of σ^2 and ρ is based on this regression structure. Let s_{e_g} be the observed residual square and cross-product terms from the lower triangle of $e_g e_g'$ stacked into an $n_g(n_g + 1)/2$ column vector. Similarly let q_{1_g}, q_{2_g} and ξ_g be the corresponding vectors of elements from Q_1, Q_2 and ξ for cluster g. The OLS regression model used to estimate σ^2 and ρ is,

$$\begin{bmatrix} s_{e_1} \\ s_{e_2} \\ \vdots \\ s_{e_g} \\ \vdots \\ s_{e_G} \end{bmatrix} = \begin{bmatrix} q_{1_1} \\ q_{1_2} \\ \vdots \\ q_{1_g} \\ \vdots \\ q_{1_G} \end{bmatrix} \sigma^2 + \begin{bmatrix} q_{2_1} \\ q_{2_2} \\ \vdots \\ q_{2_g} \\ \vdots \\ q_{2_G} \end{bmatrix} \rho + \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_g \\ \vdots \\ \xi_G \end{bmatrix}. \tag{10}$$

This gives

$$\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\rho} \end{bmatrix} = \begin{bmatrix} q_1' q_1 & q_1' q_2 \\ q_2' q_1 & q_2' q_2 \end{bmatrix}^{-1} \begin{bmatrix} q_1' s_e \\ q_2' s_e \end{bmatrix}. \tag{11}$$

This is not a statistical regression, but a method for choosing the values for σ^2 and ρ that best fit the observed squares and cross-products of the residuals, where best is defined as the minimum sum of squared errors. This method is referred to as cluster estimated standard errors (CESE).

One adjustment is made computing the CESE. Davidson and MacKinnon (1993, p. 554) recommend adjusting the residuals by $hc2 = e_i/\sqrt{1-h_{ii}}$ or by $hc3 = e_i/(1-h_{ii})$, where h_{ii} denotes the ith diagonal element in the projection matrix P_g. They argue that squared residuals underestimate the true variance of the stochastic terms, particularly for observations that exert leverage on the OLS estimates (indicated by larger values for h_{ii}). They recommend the $hc3$ adjustment for data where heteroskedasticity is present. Long and Ervin (2000) report the results of extensive Monte Carlo simulations that support the Davidson and MacKinnon recommendations. These arguments and demonstrations are the basis for these adjustments.[7] These are referred to as the $CESE_2$ and $CESE_3$ estimators, respectively.

3 Monte Carlo Experiments Comparing Correction Methods

Monte Carlo experiments are conducted to compare the performance of the two methods for estimating standard errors with clustered data—cluster robust standard errors (CRSE) and cluster estimated standard errors (CESE). The distributions of the explanatory variables and stochastic terms are selected to examine the performance of the two estimators under a variety of circumstances. The initial simulations match the ideal conditions for the CRSE—homogeneously distributed explanatory variables with no covariate clustering, an equal number of observations per cluster and homogeneous normally distributed stochastic terms. Subsequent simulations vary the number of observations per cluster, the amount of covariate clustering, the heterogeneity of the stochastic term distributions for each cluster and the shape of the distribution from which the stochastic terms are obtained. Each simulation is done for varying numbers of clusters.

The model has three explanatory variables and one interaction term,

$$Y_{g_i} = 2 + 1 * X_{1_{g_i}} + 0 * X_{2_{g_i}} + 0 * (X_{1_{g_i}} \cdot X_{2_{g_i}}) + 0.3 * X_{3_{g_i}} + V_{g_i}. \tag{12}$$

The values for β_2 and β_3 equal zero to enable calculation and comparison of the rejection rates for Wald tests of the null hypothesis that both are zero. The values for the explanatory variables are random draws from a Chi-squared distribution with three degrees of freedom. All experiments are done with 12, 24, 48 and 72 and for the ideal conditions with 96 clusters to provide a better examination of the asymptotic properties.

The estimators' performance is compared in two ways. The first is the mean standardized error in the estimated standard error of the coefficients. Let s_b be the standard deviation of the distribution of simulated coefficients and $\bar{\hat{s}}_b$ be the mean estimated coefficient standard error. The mean standardized error is mste $= (\bar{\hat{s}}_b - s_b)/s_b$.[8] This is the mean error in estimating the coefficient standard error standardized by the standard deviation of the coefficient distribution. For example, if $s_b = 1$ and the mean estimated standard error is 0.9 then mste $= -0.1$. This term is then averaged for the five coefficients in the model to give the average mean standardized error (amse). The second comparison is the percent of the simulations in which the Wald test of the null

7 The same adjustments can be done computing $(e_g e_g')$ with CRSE, Bell and McCaffrey (2002). Some of the simulations that proved most troublesome for CRSE were redone with these adjustments, with minimal improvements. These corrections are eschewed in the Monte Carlo simulations because of the small gains but more importantly so the simulated CRSE are the Liang and Zeger (1986) estimator and so the results are comparable to those with standard estimation packages. The CRSE computed here do include a degrees of freedom adjustment, $\left(\frac{G}{G-1}\right)\left(\frac{N-1}{N-K}\right)$, that is part of the Stata *regress* package. Stata also treats the t-statistic as having $G-1$ degrees of freedom in calculating the p-value.

8 Let $\hat{\sigma}_{b_r} - s_b$ be the error in the estimated coefficient standard error in the rth replication of the simulation with R total trials. Then mste $= \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\sigma}_{b_r} - s_b}{s_b}\right)$.

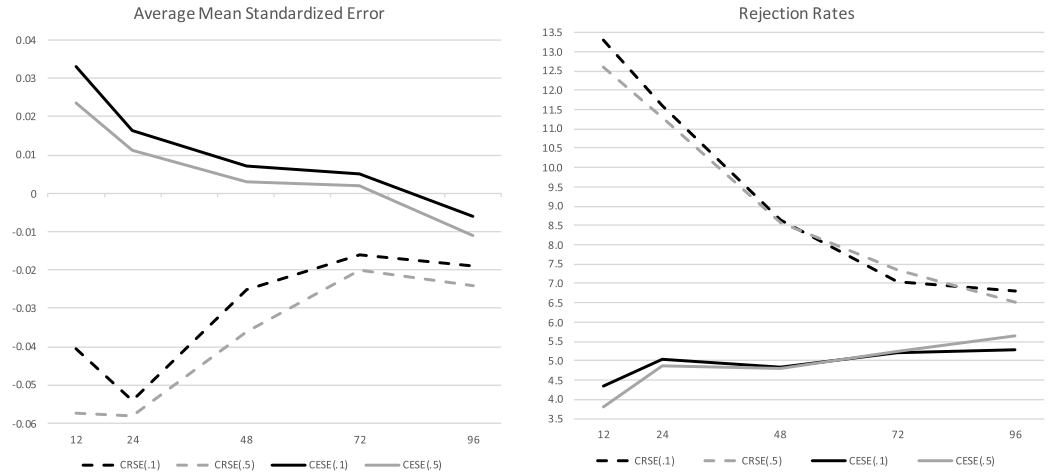

Figure 1. Estimator performance with homogeneous clusters.

hypothesis that $\beta_2 = \beta_3 = 0$ is rejected for $p = 0.05$.[9] Any value greater than 5% indicates this null hypothesis is being rejected at too high a rate, increasing the probability of a Type I error.

A complication arose in a small fraction of the simulations where the CESE estimated value for ρ is greater than the estimated value for σ_v^2. This situation occurred in 0.22% of all the simulations and was concentrated in the most problematic scenarios. Over 40% of the occurrences are in the simulations with twelve clusters and heteroskedastic error terms drawn from exponential or chi-squared distributions or with highly correlated errors, which are less than 8% of all the simulations. An arbitrary strategy is used which resets $\hat{\sigma}^2 = (\hat{\rho} + 0.02)$ whenever $\hat{\rho} > \hat{\sigma}_u^2$.[10]

3.1 Homogeneous Clusters

The initial experiments constitute the ideal sample for CRSE. There are ten observations per cluster, the explanatory variables for each cluster are drawn from distributions with the same mean and variance, and the stochastic terms are drawn from identical normal distributions for each observation and each cluster. For these experiments the number of clusters is expanded to ninety-six to observe the asymptotic properties better. The within cluster stochastic term covariance is created by specifying a cluster specific stochastic term that is included in each observation's stochastic term, $V_{g_i} = u_g + \epsilon_{g_i}$. The terms u and ϵ are independent so the stochastic term variance is $\sigma_v^2 = \sigma_u^2 + \sigma_\epsilon^2$ and the covariance is $\rho = \sigma_u^2$. For these simulations both stochastic terms are independently drawn from standard normal distributions. Let $r_v = \sigma_u^2/\sigma_v^2$ be the expected correlation among the stochastic terms in each cluster. The simulations are done with both a very low and a moderate expected correlation, $r_v = 0.1$ and $r_v = 0.5$, respectively. The two panels in Figure 1 plot the performance measures for each estimator. The number in parentheses in the legend denotes the correlation of the stochastic terms within each cluster. Online appendix Table C.1 reports the numerical results.

The CRSE asymptotic properties are very evident. The underestimates of the coefficient standard deviations range from about 5% with twelve clusters, to about 3% with forty-eight clusters to about 2% with ninety-six clusters. Similarly, the rejection rates drop from 13% with twelve clusters to about 6.5% with ninety-six clusters. The important result is that with forty-eight or fewer clusters, at least with the simulated data here, CRSE are quite unreliable and this unreliability increases sharply as the number of clusters decreases.

9 The replication programs also include the rejection rate for $p = 0.01$.

10 Greene (2012, p. 375) reports a similar problem and proposes a similar solution, such as setting a negative variance estimate to zero, when estimating comparable terms in the context of pooled time-series, cross-section models.

The CESE perform quite well for most numbers of clusters.[11] CESE overestimate the coefficient standard deviations by 2 to 3% with twelve clusters and underestimate them by less than 1% with ninety-six clusters. The CESE consistently reject the null hypothesis by close to the desired 5%. The CESE inferences are actually too conservative. In half the simulations the average rejection rate is less than 5%.

Differences with variations in the correlation of the stochastic terms in each cluster are quite small. Exceptions are the average mean standardized errors with twelve clusters, where the CRSE amse are 0.015 and the CESE amse are about 0.01 more negative with the higher correlation. For the CRSE and the CESE the differences in the rejection rates with the different amounts of interdependence are less than 0.4% except for the CRSE with twelve clusters, where the difference is about 0.7%. The subsequent question is whether cluster heterogeneity leads to larger differences with increased interdependence.

Summarizing, CRSE performed as expected with ideal data—poorly with only a few clusters and better as the number increased. The surprise is the apparent number of clusters for the asymptotic properties to dominate. MacKinnon and Webb (2017, p. 234) cite Angrist and Pischke (2009) suggesting a "rule of 42," that forty-two clusters are sufficient for reliable inference, but then argue there is no reliable rule of thumb on the necessary number of groups as the performance of CRSE is very sample specific. These simulations provide a further illustration of the MacKinnon and Webb concerns. The CESE are far less sensitive to the number of clusters, with relatively little change in performance particularly for rejection rates.

3.2 Estimator Performance with Heterogeneity

The next simulations move away from the ideal, homogeneous, sample with normally distributed stochastic terms in several important ways. These moves more closely approximate the situations encountered in Political Science. In these simulations stochastic terms are drawn from either a normal or an exponential distribution with three different types of heterogeneity.

(1) Number of observations per cluster. In these experiments there are five, ten or fifteen observations per cluster in equal proportions. This gives an average number of observations equal to that in the homogeneous experiments but with important variation.
(2) Heterogeneous explanatory variables. The heterogeneity explored here is the amount of covariate clustering, defined as the between cluster variance as a proportion of the variable's total variance. When covariate clustering equals zero the explanatory variables in each cluster have the same mean and variance. The amount of covariate clustering equals 0, 0.3, 0.6, and 0.9.[12] The higher covariate clustering corresponds to analyses where factors such as institutions or electoral rules have very little within cluster variation.
(3) Heterogeneous stochastic terms. The stochastic terms are heteroskedastic between clusters, but homoskedastic within clusters.

All the comparisons begin with sixteen basic simulations—four with different numbers of clusters by four with different amounts of covariate clustering. These sixteen simulations are done with identically normally distributed stochastic terms and then with heteroskedastic exponentially distributed stochastic terms. In the heteroskedastic simulations the standard deviations for each of the stochastic components for each cluster, σ_{g_u} and σ_{g_ε}, are drawn from

11 The CESE results use the hc_2 adjustment given the homogeneity of the stochastic terms.
12 The variations in covariate clustering are achieved by specifying that each explanatory variable has a cluster specific component and a unique component, e.g. $x_{ig} = \rho_1 u_g + \rho_2 e_i$ where $\rho_2 = \sqrt{1-\rho_1^2}$ and with ρ_1 varying from 0.00 to 0.95. The total variation in x will be one and ρ_1^2 indicates the expected between group variance as a proportion of this total variance.

a uniform distribution on the interval 0.1–2.0, giving variances that range from 0.01 to 4.0.[13] The intention is to create a maximal amount of heterogeneity. The simulations with zero covariate clustering and homogeneous normally distributed stochastic terms only differ from the previous simulations in that there are now an unequal number of observations per cluster, so comparing the results illustrates the effect of unequal numbers of observations per cluster.

The simulation designs keep the distribution of the number of observations per cluster constant as the number of clusters increases. As a consequence the total sample size increases as the number of clusters increases, $N = 10 * G$. Equation (B 4) implies that CRSE performance varies with G but not with N, which is consistent with MacKinnon and Webb (2017, footnote 3, p. 237) and Esarey and Menger (2019). An online appendix explores this implication with a series of simulations with $G = 12$ but $N = 480$ for the case with heterogeneous clusters and homoskedastic normally distributed stochastic terms. The results are consistent with expectations as all the methods' performance are nearly identical with the different sample sizes but equal numbers of clusters.

3.2.1 *Estimator Performance with Heteroskedastic and Exponentially Distributed Errors*

These simulations compare the estimators' performance in extremely unfavorable conditions—exponentially and heteroskedastically distributed stochastic terms—with that in better conditions —normally and homoskedastically distributed errors. The methods should perform worst in the former simulations and best in the latter. This provides a good picture of the range of the estimators' performance and whether there is any commonality in that performance in quite different conditions. The emphasis here is a comparison of CRSE and CESE under these two conditions. The next section has a more detailed discussion of CESE performance in a range of conditions.

The left panel in Figure 2 plots the amse and rejection rates for the results for both estimators under the more favorable conditions. The right panels show these plots for heteroskedastically, exponentially distributed errors.[14] (All the results are reported in online appendix Table A.2.) In the most favorable case, with seventy-two clusters, no covariate clustering and homoskedastic normal error terms the CRSE underestimate the true standard deviations by 2% and reject the null hypothesis of no association 7% of the time. This performance is slightly worse than what is shown in Figure 1 when there are identical numbers of observations per cluster.

The CRSE performance declines as covariate clustering increases, even with a large number of clusters, and as the number of clusters decreases, even with no covariate clustering. This strong negative interaction between these conditions is predicted by Harden's (2011) deflation factor shown in Footnote 6. The amse decreases to −0.09 and the rejection rate increases to 12% with seventy-two clusters but very high covariate clustering, to −0.08 and 15% respectively with twelve clusters and no covariate clustering, and to −0.40 and an over 50% rejection rate with twelve clusters and very high covariate clustering. The performance falls between these values for intermediate numbers of clusters and covariate clustering.

The simulations with heteroskedastic and exponentially distributed stochastic terms, the most extreme conditions, shown in the right panels of Figure 2, present a similar pattern, but with an interesting paradox for CRSE. CRSE performance decreases sharply with increased covariate clustering and with decreasing numbers of clusters and again with a strong negative interaction between the two. The paradox is that the amse are worse but the rejection rates are lower by about 2% with the more extreme stochastic term distribution. The rejection rates are still too high for

13 The precise stata commands are: $\sigma_{u_g} = 0.1 + 1.9 * \text{runiform}()$ and $\sigma_{\epsilon_g} = 0.1 + 1.9 * \text{runiform}()$. Then $U_g = \text{rnormal}(0, \sigma_{u_g})$ and $\epsilon_{g_i} = \text{rnormal}(0, \sigma_{\epsilon_g})$ for the normally distributed simulations and $U_g = \sigma_{u_g} * [\text{exponential}(1) - 1]$ and $\epsilon_{g_i} = \sigma_{\epsilon_g} * [\text{exponential}(1) - 1]$ in the exponential simulations.

14 The CESE_2 adjustment is used with the homoskedastic stochastic terms in the left side panels and the CESE_3 adjustment with the heteroskedastic stochastic terms in the right side panels.

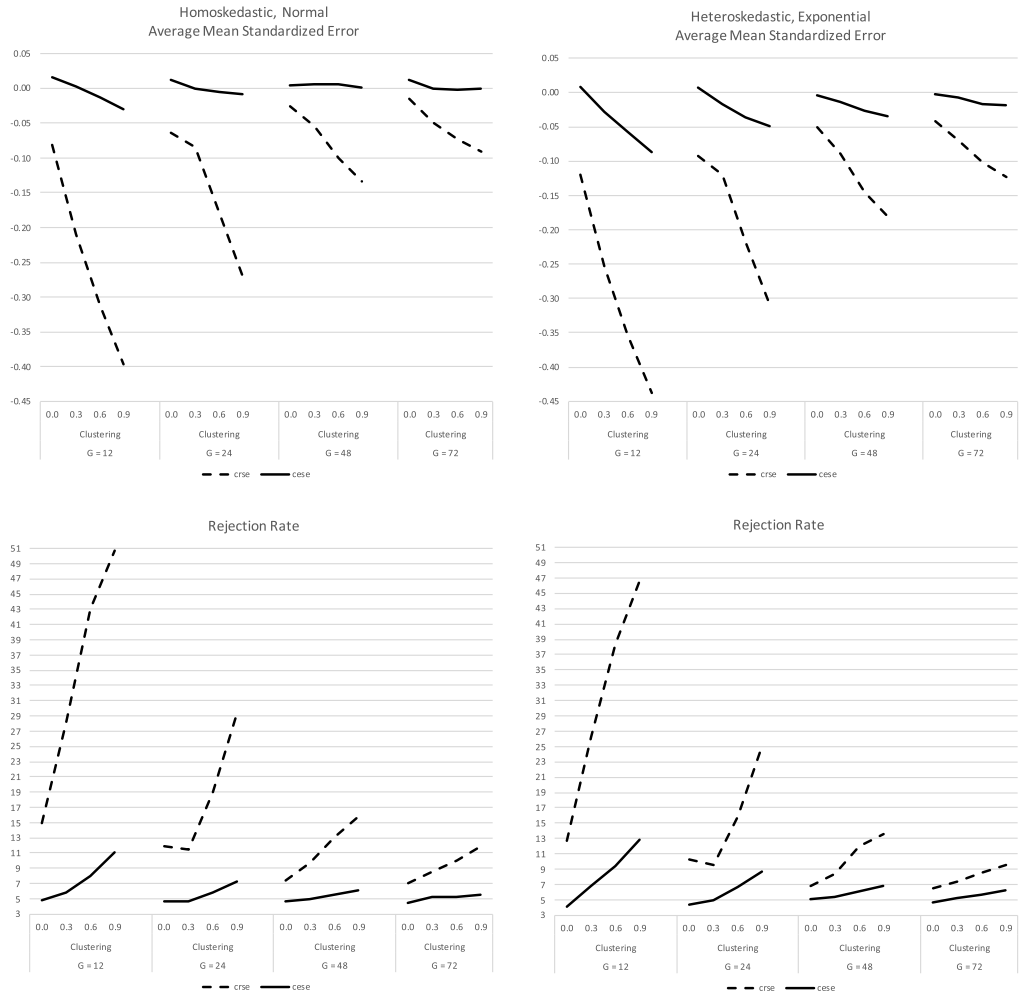

Figure 2. Estimator performance with heterogeneous clusters.

reliable inferences for all but the most favorable conditions, e.g. forty-eight or more clusters and no or a small amount of covariate clustering, where the rejection rates remain about 7%.

The CESE performance shown in Figure 2 provides a clear picture of the improvements offered by this estimator. In the best case with forty-eight or seventy-two clusters and no or a small amount of covariate clustering the CESE amse is about 0.07 smaller than with CRSE and the rejection rates are about the desired 5% compared to 7 or 8% for CRSE. The differences between CESE and CRSE increase rapidly as the number of clusters decreases and the amount of covariate clustering increases. For example, with twenty-four clusters and a moderate amount of covariate clustering the CRSE amse are about −0.20 with rejection rates of 16%. The CESE amse range from −0.01 to −0.04 and rejection rates range from 6% to 7% depending upon the stochastic term distribution. Given these consistent disparities the remaining discussion focuses on the performance of the CESE in a wide range of circumstances.

3.2.2 *CESE with Different Stochastic Term Distributions*

This section compares CESE performance across a broad range of stochastic term distributions to examine its robustness to deviations from the case with homoskedastic normally distributed errors. Stochastic terms drawn from a chi-square distribution with four degrees of freedom are added to the above distributions. These new simulations examine CESE performance with a skewed distribution but one not as extreme as the exponential. Figure 3 plots the

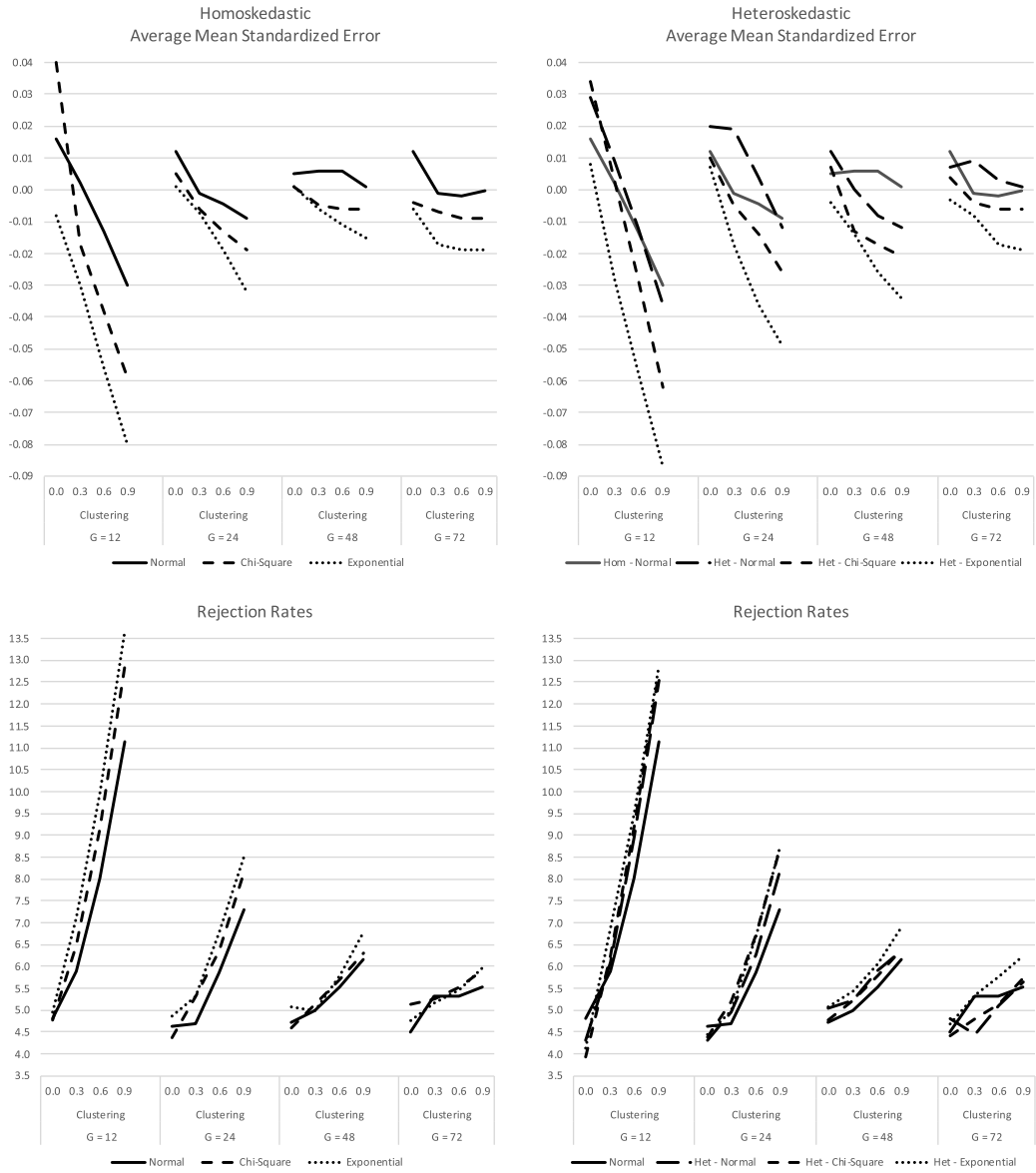

Figure 3. CESE estimator performance with different error term distributions.

CESE performance in the ninety-six different simulations. (The numerical results are shown in online Appendix A.3.) The figure highlights the performance differences associated with the different distributions, which are plotted as separate lines. The right hand plot, which shows the performance with heteroskedastic errors also includes the plots with homoskedastic normal errors, which is treated as the base case.

CESE perform quite well with a large number of clusters. With forty-eight and seventy-two clusters nine of forty-eight simulations had amse < -0.015 and eight had rejection rates greater than 6%. All nine of the cases with large amses occurred with heterogeneous non-normally distributed errors and moderate to high amounts of covariate clustering and seven of the eight cases with high rejection rates occurred with the highest level of covariate clustering, which are the most problematic scenarios. Performance declines markedly as the number of clusters decreases, particularly with only twelve clusters. With twelve clusters half the simulations have an amse < -0.02 and rejection rates greater than 7%. Again, the poor performance is concentrated in the simulations with higher levels of covariate clustering.

Low levels of covariate clustering are also associated with good performance. In the simulations with no or low amounts of clustering only five have amse < -0.015 and six have rejection rates higher than 5.5%. Increasing covariate clustering led to poorer performance. With the highest level of clustering nearly half have amse < -0.02 and half have rejection rates greater than 7%. These problems occur, with one exception, in experiments with twelve and twenty-four clusters.

Comparing performance with both covariate clustering and numbers of clusters makes the interaction between the two properties quite clear. As is evident in Figure 3 combining both a small number of clusters with high covariate clustering seriously degrades CESE performance. This combination of circumstances also increases the effects of having non-normal stochastic term distributions. In the experiments with twelve or twenty-four clusters and a moderate or high level of covariate clustering going from a normal distribution to an exponential increases the amse from -0.014 to -0.052 and the rejection rate from 8.53 to 9.59. With forty-eight or seventy-two clusters and no or low covariate clusters the comparable range for the amse is 0.006 to -0.007 and is 4.88 to 5.07 for rejection rates. Comparing simulations with homoskedastic and heteroskedastic distributions showed very little difference or interaction effects.

The CESE method performs very well, even too conservatively at times, with large numbers of clusters, small amounts of covariate clustering, or homoskedastic normally distributed stochastic terms. This performance declines with each step away from these desirable conditions, with substantial negative interactions between the number of clusters and the amount of covariate clustering. Figure 3 shows that performance becomes unacceptable, such as rejection rates that exceed 8 or 9%, with extreme deviations from the ideal.

3.2.3 *Differences in Cluster Stochastic Term Correlation*

The final simulations in this section explore how differing amounts of stochastic term correlation within a cluster affect CESE performance with heterogeneous clusters. Equation (6) shows that the amount by which CRSE underestimate the true standard deviation varies with the stochastic term correlation, ρ, times the amount of heterogeneity in the clusters. This relationship predicts that increasing ρ will degrade the performance of the estimators more in the heterogeneous than in the homogeneous case. Recall that Figure 1 shows the performance of the CRSE and CESE estimators with different amounts of interdependence, $r_v = 0.1$ and 0.5, for situations with completely homogeneous clusters—equal numbers of observations per cluster and identical distributions for the explanatory variables and the stochastic terms. We now make the same comparisons with the heterogeneous clusters used in the previous experiments with the expectation that increasing interdependence will decrease performance faster in the heterogeneous case. The heterogeneous sample has five, ten or fifteen observations per cluster, sets the covariate clustering level at 0.6 and uses normally homoskedastically distributed stochastic terms. The simulations are done for a low correlation, $r_v = 0.1$, a moderate correlation, $r_v = 0.5$ and a high correlation, $r_v = 0.75$. The results for the simulations with homogeneous clusters are shown in online Appendix A.1 and those for the heterogeneous clusters are shown in online Appendix A.4.

Figure 4 graphs the performance associated with the levels of interdependence for the homogeneous and heterogeneous simulations. (The homogeneous simulation plots repeat Figure 1.) One conclusion, evident in Figure 1, is that differences in stochastic term correlations have little effect on the CESE with homogeneous clusters. Not so with heterogeneous clusters. The performance gap is large between the low and high correlation cases with twelve clusters, is relatively small with twenty-four clusters and virtually disappears with forty-eight or more clusters. For example, with twelve clusters the amse are 0.00, -0.01 and -0.02 and the rejection rates are 6.8%, 8.0% and 9.3% across the three correlation levels. With forty-eight clusters the amse are 0.01, 0.006 and -0.005 and the rejection rates are 5.3%, 5.5% and 5.6%. The conclusion

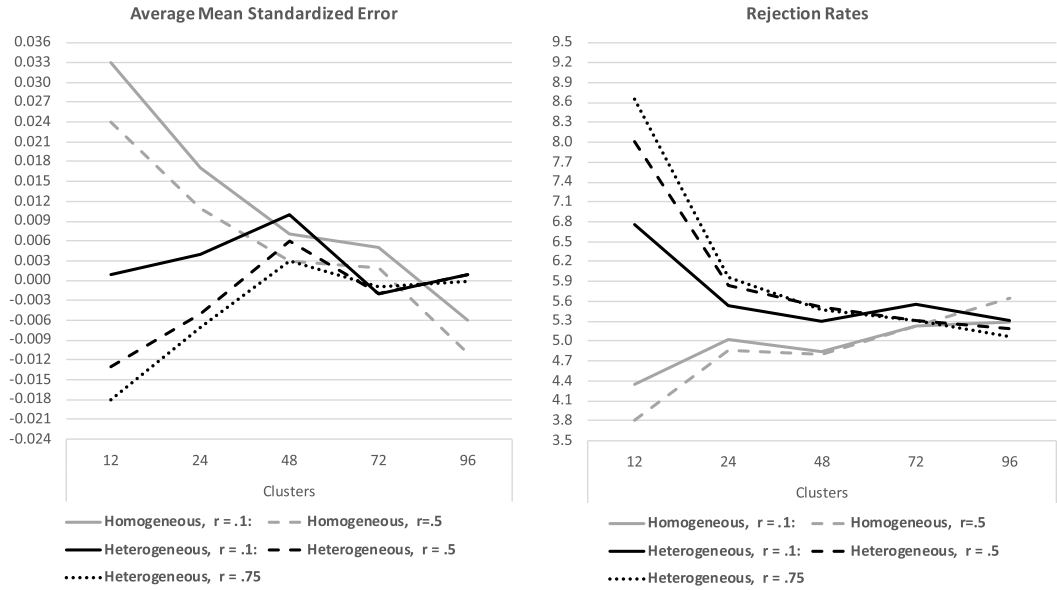

Figure 4. Estimator performance with differences in interdependence.

is that the level of interdependence matters, but primarily with low numbers of heterogeneous clusters.

3.3 Bootstrapped Standard Errors

Most of the current literature examines bootstrapping as a correction for CRSE. The bootstrapping procedure in Harden (2011) returns an estimate for Σ_b, which enables the full range of statistical tests such as the Wald tests explored in the previous simulations. This section repeats a selected set of the previous simulations using this bootstrap method to estimate coefficient standard errors and conduct the Wald tests and compares these results with the CESE. The scenarios are:

(A) homogeneous clusters with homoskedastic normally distributed stochastic terms;
(B) scenario A but with an unequal number of observations per cluster;
(C) scenario B but with covariate clustering equals 0.6;
(D) scenario C but with heteroskedastic error terms; and
(E) scenario C but with stochastic terms draws from a χ^2 distribution.

These simulations compare cluster bootstrapped standard errors (CBSE) with CESE over this range of conditions. They are not a comprehensive examination of CBSE performance. All five simulations are done with the same data as the previous simulations.[15] The scenarios are ordered in terms of increasing difficulty for the CESE estimator.

The simulations followed Harden (2011, footnote 10) in using one thousand replications for the bootstrap. The simulations raise a concern with bootstrapping. The estimated Σ_b varied noticeably with the number of replications and the random number seed. For example, the average rejection rate for the simulation with twenty-four clusters and scenario C differed by more than 1% with two different seeds. (More on this in online Appendix C.) For consistency all reported results are done with 1000 replications and the random seed used for the CESE simulations, but it is important to understand that these estimates of coefficient uncertainty are themselves uncertain.

Figure 5 plots the amse and rejection rates for the CESE and CBSE for all five scenarios with varying numbers of clusters. The quick summary is that CESE have smaller errors than the CBSE

15 The only difference between the bootstrap simulations and the previous simulations is that for computational economy the bootstrap simulations are done for 5,000 rather than 10,000 iterations.

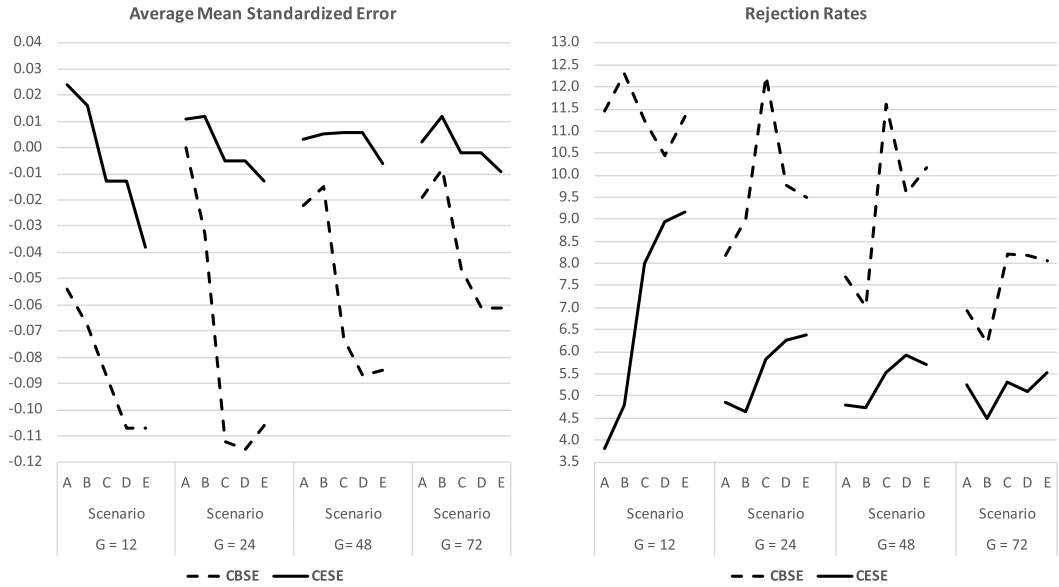

Figure 5. CESE & bootstrapped standard errors.

method for all scenarios and numbers of clusters. The CBSE amse were substantially larger for all scenarios with only twelve clusters. Above twelve clusters there are relatively small differences between the CBSE and CESE amse for the first two scenarios, which include no covariate clustering and homoskedastic normal errors. As the scenarios deviated from these more ideal conditions the difference in the two amse grows substantially. For example, with twenty-four clusters the CESE and CBSE amse are both zero for scenario A, but the CBSE amse grow almost monotonically across the scenarios, reaching −0.11 for scenarios C, D and E while the CESE amse $\geq$ −0.013. With seventy-two clusters the CBSE amse are −0.02 to −0.01 and the CESE amse are weakly positive for scenarios A and B. But, for scenarios C to E where the scenarios deviate from these ideal cases the CBSE amse decrease to −0.05 and −0.06 while all the CESE amse are −0.01 or higher.

The CBSE rejection rates follow a somewhat irregular pattern though all are higher than the CESE measures. There is very little variation in CBSE rejection rates across the simulations with twelve clusters though they are much higher than the CBSE rejection rates even in scenarios D and E, which present the greatest difficulties for CESE. With twenty-four and forty-eight clusters CBSE rejection rates increase moving from scenarios A to E with a sharp spike for scenario C. In the most favorable simulations the CBSE rejection rates are just over 6% and increase substantially after that for scenarios C, D, and E, being 10% or more in over half the fifteen simulations where $G < 72$. By comparison the CESE rejection rates are less than 6% in ten of the same fifteen simulations. With seventy-two clusters the CBSE rejection rates vary from 6.2% to 8.2% while the CESE rejection rates range from 4.5% to 5.5%.

The Monte Carlo results show that CBSE perform well in simulations with a moderate to large number of relatively homogeneous clusters, $G > 12$ and scenarios A and B. With $G = 12$ and/or heterogeneity across the clusters and in the error terms the CBSE amse become increasingly negative and the rejection rates increase beyond what should be acceptable. CESE are much less sensitive to these variations, except with $G = 12$ and heterogeneous clusters and stochastic terms, though in all scenarios they perform better than CBSE.[16]

16 A small selected set of scenarios were repeated using the wild, fast bootstrap in the Stata 'boottest' program (Roodman *et al.* 2019) specified to return the full coefficient variance–covariance matrix. The results had larger errors than the CBSE results shown above and were more sensitive to the difficulties in scenarios C, D and E.

3.4 Monte Carlo Simulation Summary

There are several take-aways from the simulations. As is now well established CRSE seriously underestimate the standard deviation of the simulated coefficient distributions and produce confidence intervals that are far too narrow, leading to rejection rates that are substantially too high except with homogeneous clusters and a large number of clusters. A common solution is cluster bootstrapped standard errors, particularly as presented in Harden (2011) where the full coefficient variance–covariance matrix is estimated. In the simulations done here this method is a vast improvement on CRSE, but still underestimates the actual coefficient standard deviations, except for data with very homogeneous clusters, and over rejects a null hypothesis using a Wald test based on the estimated coefficient variance–covariance matrix. An option proposed here is the CESE correction, which has better estimated standard errors and rejection rates. The quality of this estimator decreases as the number of clusters decreases, as the amount of covariate clustering increases, and with heterogeneous stochastic terms. In terms of relative performance in the simulations conducted here the CESE method outperformed the CBSE method particularly as the number of clusters decreases, as the amount of covariate heterogeneity increases and as the stochastic terms deviate from being homoskedastically normally distributed. These are the circumstances where CESE perform poorly, but their performance does not decline as fast or as far as the other methods examined.[17]

4 Examples

We compare the CRSE, CBSE and CESE estimators with two examples.[18] One from state politics, Brown, Jackson, and Wright (1999), which Harden (2011) uses to compare CRSE and CBSE. The other from comparative politics, Elgie *et al.* (2014). The former has relatively homogeneous clusters as each state has four observations, which should favor CRSE and CBSE. In the comparative politics example the number of observations per cluster ranges from one to thirty-two, which should present problems for CRSE and likely CBSE.

4.1 State Politics

Harden (2011) replicates Brown, Jackson, and Wright (1999)'s model of state voter registration.[19] The propositions are that liberal (Democratic) control of the legislature, ease of registration, and party competition are associated with higher levels of voter registration. Controls for education, income, residential mobility, unemployment, south, and presidential election years are included. Harden estimates their model with both CRSE and CBSE. Liberal control and ease of registration are statistically significant with CRSE along with the education and mobility controls. With CBSE both liberal party control and mobility are no longer significant. Harden [pp. 235 and 236], however, makes the point that liberal control, "... is just on the edge of significance with CBSE ($t = 1.94$)." and "... the authors have good reason to interpret a t-value of 1.94 as support for their hypothesis."

The registration equation is replicated using CRSE, CBSE and CESE.[20] Table 1 displays the respective coefficients, standard errors and p-values.[21] These data should favor CRSE and CBSE as there are an equal number of observations per cluster. The results replicate Harden's findings

17 An online appendix examines an alternative method, which is inferior to CESE, particularly with decreasing numbers of clusters, but is superior to the other estimators. Both procedures show that estimating Σ_g, which is used to compute Σ_b, is better than the alternatives.

18 The examples are done in R using the packages lmtest, rms and ceser. ceser is available at devtools::install_github("DiogoFerrari/ceser").

19 I want to thank Professor Harden for sharing these data and software. They are exactly what a replication dataset should be, enabling both replication and extensions.

20 Online Appendix C shows the bootstrapping results vary substantially with the random number seed and the number of replications. Here the random seed is 441,022 with 50,000 replications.

21 CRSE and CESE compute p-values based on the degrees of freedom adjustments in footnote 7. The CBSE p-values are those reported by the bootcov package.

Table 1. State voter registration rates.

		St. errors			p-values		
Variable	**Coeff**	**CRSE**	**CBSE**	**CESE**	**CRSE**	**CBSE**	**CESE**
Liberal control	1.64	0.736	0.838	0.841	0.031	0.052	0.057
Registration ease	0.06	0.020	0.023	0.027	0.006	0.013	0.031
Party comp.	−0.03	0.082	0.091	0.084	0.677	0.709	0.685
Education	0.40	0.152	0.195	0.204	0.012	0.043	0.057
Income	−0.15	0.316	0.333	0.367	0.644	0.660	0.691
Unemployment	0.72	0.431	0.463	0.457	0.101	0.120	0.121
Mobility	−0.56	0.245	0.293	0.330	0.026	0.056	0.096
South	4.02	2.521	3.023	3.166	0.118	0.185	0.210
Pres. year	3.64	0.689	0.685	0.657	0.000	0.000	0.000
Constant	45.53	11.490	13.131	13.974	0.000	0.001	0.002

with CBSE standard errors being almost 15% larger than the CRSE standard errors. The CESE standard errors are, on average, 5% larger than the CBSE standard errors. Consistent with the simulations, CESE attach the most uncertainty to the coefficient estimates and CRSE the least with the uncertainty implied by CBSE being much closer to that with CESE.

This increasing uncertainty has implications for those who persist in reporting statistical significance. With CRSE a 95% confidence interval for the Liberal control, registration ease, education and mobility coefficients excludes zero. With CBSE this list is reduced to registration ease and education and with CESE only the registration ease coefficient meets this criteria.

4.2 Comparative Politics

The second example provides an important contrast to the state politics example as there are a wide range of observations per cluster making the clusters very heterogeneous. It also includes an interaction term so there is a need to use a joint test of a pair of coefficients. The example is a slightly modified version of a model relating the effective number of parties to the number of presidential candidates and presidential power presented by Elgie *et al.* (2014) building on work of Golder (2006) and Hicken and Stoll (2012).[22] They relate the number of effective legislative parties to the number of presidential candidates, a measure of presidential power, the proximity of presidential and legislative elections, the effective number of ethnic groups and the log of average district magnitudes and two interaction terms. The model the authors say shows the strongest relationship between the effective number of legislative parties and presidential power and the effective number of candidates is shown in a marginal effects plot in their Figure 4 and included in the replication file but the coefficients and standard errors are not shown in the table of results.[23]

Their Figure 4 model is re-estimated with three changes that expand the sample and number of countries. The databases developed by Golder and colleagues (Golder 2005; Bormann and Golder 2013), which form the basis for most of this comparative research, contain data that permit replacement of missing values in the proximity and log district magnitude variables. This adds twenty observations but does not increase the number of countries as these countries had nonmissing values for these variables. Seven missing values in the preferred measure of

22 I want to thank Professor Elgie for sharing their data and stata .do files. Again, these are the epitome of what a replication dataset should be.

23 Elgie *et al.* (2014, Table 1) show results for five different models estimated with a variety of standard error corrections, including CRSE, but for some unstated reason do not report the coefficients for this model. The values are easily obtained from the replication dataset and .do file. The variable measuring presidential power in their Figure 4 is labeled fapres3 in the replication data.

Table 2. Model for number of legislative parties.

		St. errors			p-values		
Variable	**Coeff.**	**CRSE**	**CBSE**	**CESE$_3$**	**CRSE**	**CBSE**	**CESE$_3$**
#Pres. candidates	0.30	0.233	0.307	0.341	0.206	0.331	0.386
Pres. power	−0.63	0.201	0.300	0.372	0.003	0.037	0.097
#Cand * pres. power	0.21	0.078	0.099	0.109	0.009	0.033	0.059
Proximity	0.01	0.249	0.278	0.370	0.958	0.963	0.972
#Ethnic groups	0.04	0.135	0.159	0.167	0.785	0.817	0.826
log(Av. magnitude)	−0.11	0.416	0.471	0.451	0.799	0.821	0.814
#Ethnic grps *log(Mag)	0.29	0.259	0.286	0.229	0.263	0.308	0.207
Constant	2.58	0.502	0.910	1.225	0.000	0.005	0.041
N	310						
#Countries	51						
Wald Test[a]		9.88	4.78	3.80	0.007	0.092	0.150

[a] Wald test that $\beta_2 = \beta_3 = 0$.

presidential power are replaced by the values in the alternative measure that Elgie *et al.* use in most of their analyses, which adds seven cases and two countries.[24] These changes expand the number of observations from 281 to 310 and increase the number of countries from 47 to 51.[25] The results of these changes favor their core hypothesis as the two coefficients relating the effective number of legislative parties to presidential power are slightly larger, though within sampling variability, and are individually and collectively more significant.[26]

Table 2 shows the estimated coefficients, standard errors and continuous p-values. The CESE standard errors are the largest, being about 55 and 20% larger than the CRSE and CBSE standard errors, respectively, for the four institutional variables.[27] As in the previous example, CRSE imply a much higher level of certainty about the coefficient distributions than either CBSE or CESE with the CBSE estimates closer to the CESE estimates than to the CRSE estimates. For example, the estimated relationship between the number of legislative parties and presidential power is −0.63. The estimated standard deviation of the distribution from which this value is drawn ranges from 0.20 with the CRSE to 0.37 with CESE.

The standard error differences are consequential for statistical inference for those who rely on these evaluations. A major point in Hicken and Stoll (2012) and Elgie *et al.* (2014) is a significant relationship between the number of legislative parties and presidential power and its interaction with the number of candidates. The p-value of the coefficient on presidential power increases from less than 0.01 level with the CRSE to less than the 0.05 level with the CBSE to just less than 0.10 level with CESE standard errors. The p-value for the interaction term coefficient increases from 0.009 to 0.032 to 0.059 with the CRSE, CBSE and CESE methods, respectively. At the conventional 95% confidence level these two coefficients are only significant with the first two methods.

Wald tests of the joint hypothesis that $\beta_2 = \beta_3 = 0$, meaning no association between the number of legislative parties and the presidential power variables, are 9.88, 4.77 and 3.80 with p-values of 0.007, 0.092 and 0.150 for the CRSE, CBSE and CESE estimates, respectively. The CRSE results clearly reject while the CBSE and CESE results do not reject this null hypothesis at the 95%

24 The difference in the two measures is that the preferred measure separates the highest category in the second measure into an additional category.

25 Two corrections are made to their data. Their replication data included ethnic fractionalization rather than the effective number of ethnic groups for Cape Verde, which is corrected. (The latter is the reciprocal of the former.) The number of presidential candidates in Nigeria in 1979 is reported as zero, which seems implausible. The Golder (2005) data report a value of 4.03 for Nigeria in 1979, which is substituted for the zero value.

26 The absolute differences in the two estimates for each coefficient are less than half their CRSE standard error.

27 The results are the CESE$_3$ adjusted standard errors as heteroskedastic errors are likely.

significance level. The CBSE value, however, is significant at the 0.10 level. This result combined with the CBSE tests of the individual coefficients, which reject the null hypothesis of no association for the two presidential power variables, might cause some doubt about not rejecting the null hypothesis, depending upon ones tolerance for Type I and Type II errors. The CESE results leave much less doubt about not rejecting the null hypotheses.

The Comparative Politics results parallel the State Politics results and the simulations. The CESE estimates are the most conservative and the CRSE estimates the least conservative. Further, the CBSE and CESE are more consistent with each other and very different from the CRSE. The important difference between the CBSE and CESE results is that the CBSE results reject the null hypothesis that the individual coefficients on the two presidential power variables, including the #Candidates*Pres. Power interaction term, are zero while the CESE estimates do not.

5 Concluding Remarks

The comparisons shown in Tables 1 and 2 and their substantive implications along with the analytical discussions and the Monte Carlo simulations indicate that the choice of an estimation method has important consequences. Resort to CRSE when analyzing grouped data is unwise, particularly with a small number of clusters, highly clustered explanatory variables or non-normally distributed errors. The sensitivity of CRSE to covariate clustering is seldom discussed in the literature, except for Harden (2011), but for many Political Science applications this will be a critical factor as many of the important variables have little variation within clusters. A method such as CESE that estimates the within cluster variance–covariance matrix and uses that estimate to compute the standard errors performs much better in simulations and in the examples is far less willing to reject a null hypothesis of no association than are methods based on the sandwich estimator associated with Liang and Zeger (1986). Bootstrap methods, commonly proposed as the alternative to CRSE, perform much better than CRSE but still have smaller estimated standard errors and higher rejection rates than CESE in both the simulations and the examples. In any specific application one may want to compare the performance of the different methods for the data at hand using the type of Monte Carlo simulations undertaken here given their likely sample specific performance.

The next challenge is to explore if, or how, the CESE estimator can be extended to the general linear model, which includes the important class of limited dependent variables, counts, etc. We lay out a possible way to begin this exploration. The discussion is restricted to the general linear model where $Y_i = f(X_i\beta) = f(z_i)$ and the log likelihood function for observation i is $L_i^* = g(z_i)$. Let $\delta_i = \partial L_i^*/\partial z_i$ so that $\partial L_i^*/\partial \beta_k = \delta_i X_{ik}$. The MLE estimates for β are obtained by solving the following set of equations, $\partial L^*/\partial \beta_k = \sum_{i=1}^{N} \delta_i \beta_k = 0, k = 1, \ldots, K$. Further, let D equal the inverse of the Hessian matrix of second derivatives, $D = [\partial^2 L^*/\partial \beta_k \partial \beta_{k'}]^{-1}$. The sandwich estimate for the variance–covariance matrix of the MLE coefficients is

$$V(\hat{\beta}) = D^{-1}[X'(\delta\delta')X]D^{-1}. \tag{13}$$

With the OLS model $\delta_i = (Y_i - X_i\beta) = u_i$. For the standard OLS model with iid stochastic terms $E(uu') = \sigma_u^2 I$ and Equation (13) becomes the familiar $V(\hat{\beta}) = \sigma_u^2(X'X)^{-1}$. With clustered data Equation (13) becomes Equation (1). Substituting the residuals e_i for u_i gives Equation (3), the expression for CRSE.

Equation (13) is the starting point in developing the equivalent to CESE for the general linear model. It will be necessary to find expressions for δ_i and then relate the matrix $\delta\delta'$ to the interdependence within clusters. Although these steps are not obvious for many applications the logit model for binary outcomes offers an example of how one might begin. In the logit model $L_i^* = -\log(1 + e^{-z_i})$ for $Y_i = 1$ and $-\log(1 + e^{z_i})$ for $Y_i = 0$. With P_i denoting the probability

that $Y_i = 1$ the resulting expressions for δ_i are:

$$\delta_i = e^{-z_i}/(1 + e^{-z_i}) = 1 - P_i : \quad Y_i = 1, \tag{14}$$

and

$$\delta_i = -e^{z_i}/(1 + e^{z_i}) = 0 - P_i : \quad Y_i = 0. \tag{15}$$

Fortuitously the δ_i are analogous to the stochastic terms in the OLS model as they indicate how the observed outcomes, which are measured as zeros and ones, differ from the true probability that $Y_i = 1$. If after calculating the logit estimates for β the predicted probabilities $\hat{P}_i$ are calculated the residuals $d_i = 1 - \hat{P}_i$ or $d_i = -\hat{P}_i$ can be substituted for δ in Equation (13). The next, and most difficult step, is to derive an expression for dd' as a function of the true variances and covariances within clusters, analogous to Equations (9) and (11), that can be used to estimate $\delta\delta'$ in Equation (13).

Data Availability Statement

Replication materials can be found at Jackson (2019).

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2019.38.

Appendix A. Expression for $E(e_g e_g')$

This appendix develops the expression Equation (4).

$$\begin{aligned}
E(e_g e_g') &= E[(Y_g - X_g b)(Y_g - X_g b)'] = E\{[V_g - X_g(b-\beta)][V_g - X_g(b-\beta)]'\} \\
&= E\{[V_g - X_g(X'X)^{-1}X'V][V_g - X_g(X'X)^{-1}X'V]'\} \\
&= E(V_g V_g') - E(V_g V')X(X'X)^{-1}X_g' - X_g(X'X)^{-1}X'E(VV_g') \\
&\quad + X_g(X'X)^{-1}X'E(VV')X(X'X)^{-1}X_g' \\
&= \Sigma_g - \Sigma_g P_g - P_g\Sigma_g + X_g(X'X)^{-1}\left(\sum_{g=1}^{G} X_g'\Sigma_g X_g\right)(X'X)^{-1}X_g',
\end{aligned} \tag{A 1}$$

where P_g is the projection matrix $P_g = X_g(X'X)^{-1}X_g'$. This equation makes repeated use of the condition that the stochastic terms are independent across clusters.

Equation (A 1) can be extended with repeated use of the assumption of homogeneity within clusters and the expression that $\Sigma_g = \rho\iota_g\iota_g' + (\sigma_v^2 - \rho)I_g$. (Recall that ι_g is an n_g x 1 vector of ones and I_g is an n_g x n_g identity matrix.) Use this expression to rewrite Equation (A 1) as,

$$\begin{aligned}
E(e_g e_g') &= \rho_g\iota_g\iota_g' + (\sigma_g^2 - \rho_g)I_g - 2(\sigma_g^2 - \rho_g)P_g - \rho_g(P_g\iota_g\iota_g' + \iota_g\iota_g'P_g) \\
&\quad + X_g(X'X)^{-1}\left\{\sum_{g=1}^{G}[(\sigma_g^2 - \rho_g)X_g'X_g + \rho_g X_g'\iota_g\iota_g'X_g]\right\}(X'X)^{-1}X_g' \\
&= \rho_g\iota_g\iota_g' + (\sigma_g^2 - \rho_g)I_g - 2(\sigma_g^2 - \rho_g)P_g - \rho_g(P_g\iota_g\iota_g' + \iota_g\iota_g'P_g) \\
&\quad + (\sigma_g^2 - \rho_g)X_g(X'X)^{-1}\left(\sum_{g=1}^{G} X_g'X_g\right)(X'X)^{-1}X_g' \\
&\quad + \rho_g X_g(X'X)^{-1}\left(\sum_{g=1}^{G} X_g'\iota_g\iota_g'X_g\right)(X'X)^{-1}X_g'
\end{aligned}$$

$$= \sigma_g^2(I_g - P_g) + \rho_g \left[\iota_g \iota_g' - (I_g - P_g) - (P_g \iota_g \iota_g' + \iota_g \iota_g' P_g) \right.$$
$$\left. + X_g (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \iota_g \iota_g' X_g \right) (X'X)^{-1} X_g' \right]$$
$$= \sigma_g^2 Q_1 + \rho_g Q_2, \tag{A 2}$$

where Q_1 and Q_2 are functions of the observed explanatory variables. This equation shows that $E(e_g e_g')$ is a linear function of the unknown terms σ_g^2 and ρ_g.

Appendix B. CRSE—Bivariate Case

This appendix uses the expression for $E(e_g e_g')$ from Equation (A 1) in the previous appendix to derive the CRSE estimate for the variance of the estimate in the bivariate case. In this derivation all variables are mean centered, e.g. $x_{gn} = (X_{gn} - \bar{X})$, removing the constant term. The bivariate case along with mean centered variables greatly simplify the algebra because the term $(X'X)^{-1}$ reduces to the scalar $1/NV_x$ where N is the total sample size and V_x is the full sample variance of the explanatory variable. We also simplify the model by assuming homogeneous error terms, $\Sigma_{vg} = \Sigma_v$ for all g.

Begin with the expression for $\sum_{g=1}^{G} x_g' \Sigma_v x_g$.

$$\sum_{g=1}^{G} x_g' \Sigma_v x_g = \sum_{g=1}^{G} (x_{g1}, x_{g2}, \ldots, x_{gn_g}) \begin{pmatrix} \sigma^2 & \rho & \ldots & \rho \\ \rho & \sigma^2 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & \sigma^2 \end{pmatrix} \begin{pmatrix} x_{g1} \\ x_{g2} \\ \vdots \\ x_{gn_g} \end{pmatrix}$$
$$= \sum_{g=1}^{G} \left[\sigma^2 (x_{g1}^2 + x_{g2}^2 + \cdots + x_{gn_g}^2) + 2\rho \left(\sum_{i=1}^{n_g - 1} \sum_{j=i+1}^{n_g} X_{gi} X_{gj} \right) \right]$$
$$= \sigma^2 \sum_{g=1}^{G} \sum_{i=1}^{n_g} X_{gi}^2 + 2\rho \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g - 1} \sum_{j=i+1}^{n_g} C_{g_{ij}} \right)$$
$$= \sigma^2 N V_x + 2\rho \left(\sum_{g=1}^{G} C_g^* \right), \tag{B 1}$$

where $C_{g_{ij}}$ is the cross-product of the values for x_g in periods i and j and C_g^* is the sum of these terms in cluster g. Equation (B 1) is an important term as it appears in the equation for the true coefficient variance–covariance matrix, Equation (1), and is the first term and the sandwich term in large parentheses in the expression for the expected coefficient variance–covariance matrix using CRSE, obtained from substituting Equation (4) in Equation (3).

The next term to derive is $W = \sum_g x_g' \Sigma_v P_g x_g = \sum_g x_g' \Sigma_v [x_g (x'x)^{-1} x_g'] x_g$.

$$W = \sum_{g=1}^{G} [(x_g' \Sigma_v x_g)(x'x)^{-1} (x_g' x_g)]$$
$$= \frac{1}{NV_x} \sum_{g=1}^{G} \{ [\sigma^2 (x_{g1}^2 + \cdots + x_{gn_g}^2) + 2\rho C_g^*] (x_{g1}^2 + \cdots + x_{gn_g}^2) \}$$
$$= \frac{\sigma^2}{NV_x} \sum_{g=1}^{G} (n_g V_{x_g})^2 + \frac{2\rho}{NV_x} \sum_{g=1}^{G} n_g V_{x_g} C_g^*, \tag{B 2}$$

where V_{x_g} is the variation of x about the full sample mean in cluster g. Given the symmetries the value for $x_g' P_g \Sigma x_g$ will be identical to W in Equation (B 2).

Last there is an expression for $Z = \sum_g x_g'[x_g(x'x)^{-1}(\sum_g x_g'\Sigma_v x_g)(x'x)^{-1}x_g']x_g$.

$$\begin{aligned} Z &= \frac{1}{(NV_x)^2}\sum_{g=1}^{G}[(x_{g1}^2 + \cdots + x_{gn}^2)^2(\sigma^2 NV_x + 2\rho C_g^*)] \\ &= \frac{\sigma^2}{NV_x}\sum_{g=1}^{G}(n_g V_{x_g})^2 + \frac{2\rho}{(NV_x)^2}\sum_{g=1}^{G}(n_g V_{x_g})^2 C_g^*. \end{aligned} \tag{B 3}$$

Substituting the expressions in Equations (B 1)–(B 3) for the terms in Equation (3) gives the expected value of the CRSE estimated coefficient variance as,

$$\begin{aligned} E(S_b^2) &= \frac{1}{(NV_x)^2}\left[\sum_{g=1}^{G}(x_g'\Sigma_v x_g) - 2W + Z\right] \\ &= \frac{1}{(NV_x)^2}\left\{\sigma^2 NV_x + 2\rho\sum_{g=1}^{G}C_g^* - \frac{2}{NV_x}\left[\sigma^2\sum_{g=1}^{G}(n_g V_{x_g})^2 + 2\rho\sum_{g=1}^{G}n_g V_{x_g}C_g^*\right]\right. \\ &\quad \left. + \left[\frac{\sigma^2}{NV_x}\sum_{g=1}^{G}(n_g V_{x_g})^2 + \frac{2\rho}{(NV_x)^2}\sum_{g=1}^{G}(n_g V_{x_g})^2 C_g^*\right]\right\} \\ &= \frac{\sigma^2}{NV_x}\left[1 - \frac{\sum_{g=1}^{G}(n_g V_{x_g})^2}{(NV_x)^2}\right] + \frac{2\rho}{(NV_x)^2}\left\{\sum_{g=1}^{G}\left[1 - \frac{2n_g V_{x_g}}{NV_x} + \frac{(n_g V_{x_g})^2}{(NV_x)^2}\right]C_g^*\right\} \\ &= \frac{\sigma^2}{NV_x}\left[1 - \sum_{g=1}^{G}\left(\frac{n_g V_{x_g}}{NV_x}\right)^2\right] + \frac{2\rho}{(NV_x)^2}\left[\sum_{g=1}^{G}\left(1 - \frac{n_g V_{x_g}}{NV_x}\right)^2 C_g^*\right]. \end{aligned} \tag{B 4}$$

From Equations (1) and (B 1) the true coefficient variance is,

$$\sigma_b^2 = (x'x)^{-1}\left(\sum_{g=1}^{G}x_g\Sigma_v x_g\right)(x'x)^{-1} = \frac{\sigma^2}{NV_x} + \frac{2\rho}{(NV_x)^2}\sum_{g=1}^{G}C_g^*. \tag{B 5}$$

The difference between Equations (B 4) and (B 5) is the presence of the squared terms involving $n_g V_{x_g}/NV_x$, which is the share of the total variance in X contributed by cluster g. The larger these terms the larger the amount by which the CRSE underestimate the true standard errors.

B.1 CRSE with Fully Homogeneous Clusters

A sufficient condition for consistent CRSE is fully homogeneous clusters, by which we mean an equal number of observations in each cluster and that X has the same mean and variance in each cluster; $n_g = n$, $\bar{X}_g = \bar{X}$, and $V_{x_g} = V_x$ for all g. With these conditions,

$$\frac{n_g V_{x_g}}{NV_x} = \frac{nV_x}{nGV_x} = \frac{1}{G} \tag{B 6}$$

and equation (6) becomes,

$$E(S_b^2) = \frac{\sigma^2}{NV_x}\left(1 - \frac{1}{G}\right) + \frac{2\rho}{(NV_x)^2}\left[\sum_{g=1}^{G}\left(1 - \frac{1}{G}\right)^2 C_g^*\right]. \tag{B 7}$$

Thus, $\lim_{G\to\infty} E(S_b^2) = \sigma_b^2$.

References

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bell, R. M., and D. F. McCaffery. 2002. "Bias Reduction and Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 26(2):169–181.

Bormann, N.-C., and M. Golder. 2013. "Democratic Electoral Systems Around the World, 1946–2011." *Electoral Studies* 32:360–369.

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

Cameron, C. A., J. B. Gelbach, and D. L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3):414–427.

Davidson, R., and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York, NY: Oxford University Press.

Eicker, F. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. Le Cam and J. Heyman, 59–82. Berkeley, CA: California University Press.

Elgie, R., C. Bueur, B. Dolez, and A. Laurent. 2014. "Proximity, Candidates, and Presidential Power: How Directly Elected Presidents Shape the Legislative Party System." *Political Research Quarterly* 67(3):467–477.

Esarey, J., and A. Menger. 2019. "Practical and Effective Approaches to Dealing with Clustered Data." *Political Science Research and Methods* 7(3):541–559.

Franzese, R. J. Jr.. "Empirical Strategies for Various Manifestations of Multilevel Data." *Political Analysis* 13(4):430–446.

Golder, M. 2005. "Democratic Electoral Systems Around the World, 1946–2000." *Electoral Studies* 24:103–121.

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

Greene, W. H. 2012. *Econometric Analysis*. Upper Saddle River, NJ: Prentice-Hall.

Harden, J. J. 2011. "A Bootstrap Method for Conducting Statistical Inference with Clustered Data." *State Politics and Policy Quarterly* 11(2):223–246.

Hicken, A., and H. Stoll. 2012. "Are All Presidents Created Equal? Presidential Powers and the Shadow of Presidential Elections." *Comparative Political Studies* 46(3):291–319.

Huber, P. J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. Le Cam and J. Heyman, 221–223. Berkeley, CA: California University Press.

Ibragimov, R., and U. K. Muller. 2002. "t-Statistic Based Correlation and Heterogeneity Robust Inference." *Journal of Business and Economic Statistics* 28(4):453–468.

Imbens, G. W., and M. Kolesár. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *The Review of Economics and Statistics* 98(4):701–712.

Jackson, J. E. 2019 "Replication Data for: Corrected Standard Errors with Clustered Data." https://doi.org/10.7910/DVN/IABJEB, Harvard Dataverse, V1.

Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal Data Analysis for Generalized Linear Models." *Biometrika* 73:13–22.

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

MacKinnon, J. G., and M. D. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32(2):233–254.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using Boottest." *The Stata Journal* 19(1):4–60.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar. 2019. "Moving to a World Beyond '$p < 0.05$'." *The American Statistician* 73(1):1–19.

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