The Cluster-Robust Variance-Covariance Estimator: A (Stata) Practitioner's Guide

Austin Nichols and Mark Schaffer

21 Sept 2007

The problem

Suppose we have a regression model like

$$y_{mt} = x_{mt}\beta + \nu_{mt}$$

where the indexes are

$$m = 1...M \quad t = 1...T_m$$

The ms index "groups" of observations and the ts index individual observations within groups. The t suggests multiple observations over time, but the t index can represent any arbitrary index for observations grouped along two dimensions. The m subscript in T_m denotes that we may have groups of different sizes ("unbalanced" groups). It will also be convenient to have a variable `id` that identifies groups.

We assume weak exogeneity, i.e., $E(x_{mt}\nu_{mt}) = 0$, so the OLS estimator is consistent.

However, the classical assumption that ν_{mt} is *iid* (independently and identically distributed) is clearly violated in many cases, making the classical OLS covariance estimator inconsistent.

Clustered Errors

Model:

$$y_{mt} = x_{mt}\beta + \nu_{mt}$$

A natural way of modeling the structure of the covariance of ν is to assume "clustered errors": observations within group m are correlated in some unknown way, but groups m and j do not have correlated errors.

Thus

$$E(v_{mt}\nu_{ms}) \neq 0$$

$$E(v_{mt}\nu_{js}) = 0$$

and the variance-covariance matrix of ν is block-diagonal: zero across groups, nonzero within groups.

This kind of problem arises all the time.

Examples of Clustering

Example:

We have a survey in which blocks of observations are selected randomly, but there is no reason to suppose that observations within block have uncorrelated errors. For example, consider a random sample of schools that contain students whose response to some policy X might be correlated (in which case m indexes school and t indexes student within school).

Example:

$$y_{mt} = x_{mt}\beta + u_m + e_{mt}$$

where we decompose the error $\nu_{mt} = u_m + e_{mt}$ and the e_{im} are *iid*.

This is the standard "error components" model in panel data. It is traditionally addressed using the fixed or random effects estimators. If the e_{mt} are *iid*, the standard variance-covariance estimator is consistent. But....

More Examples of Clustering

Example:

$$y_{mt} = y_{mt}\beta + u_m + e_{mt}$$

where now the the e_{mt} may be serially correlated, $E(e_{mt}e_{ms}) \neq 0$. If this is the case, the fixed or random effects estimators will be consistent, but the standard covariance estimators will not.

Example:

Observations are randomly sampled, but the explanatory variable X is measured at a higher level (see Moulton 1990; Bertrand, Duflo, and Mullainathan 2004). For example, students might be randomly sampled to model test scores as a function of school characteristics, but this will result in clustered errors at the school level. If students were randomly sampled to model test scores as a function of classes taken (measured at the individual level), but classes taken and their effects on test scores are correlated within school, clustering of errors at the higher (school) level may result.

Still More Examples of Clustering

Example: Systems of equations

Say we have a two-equation model:

$$y_{m1} = x_{m1}\beta_1 + \nu_{m1}$$

$$y_{m2} = x_{m2}\beta_2 + \nu_{m2}$$

Say that ν_{m1} and ν_{m1} are both *iid*, so that each equation could be estimated separately and the standard covariance estimator would be consistent. However, we want to test cross-equation restrictions involving both β_1 and β_2. If we "stack" the data and estimate as a (seemingly-unrelated) system, the disturbances will be "clustered": $E(\nu_{m1}\nu_{m2}) \neq 0$ because the error for the *m*th observation in equation 1 will be correlated with the error for the *m*th observation in equation 2.

Still More Examples of Clustering

Example: Spatial Autocorrelation

We have data on US counties or cities. We expect counties or cities that are geographically close to each other will share some unobservable heterogeneity, but localities that are far apart will be less correlated; that is, our data are spatially autocorrelated. However, the nature of the problem we are investigating is such that US states can be regarded as essentially independent (e.g., they run separate legal, educational and tax systems). Thus it is reasonable to assert that observations on localities are independent across states but dependent within states, i.e., they are clustered by state.

NB: This is why the number 50 is of particular interest. Thus the cluster-robust covariance estimator relies on asymptotics where the number of clusters goes off to infinity. Is 50 far enough on the way to infinity?

How to Deal with Clustered Errors?

Model: $y_{mt} = x_{mt}\beta + \nu_{mt}$

Structure of the disturbances is block-diagonal:

$$Var(\nu) = \begin{pmatrix} \Sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \Sigma_m & & \\ & & & \ddots & \\ 0 & & & & \Sigma_M \end{pmatrix}$$

Two questions: (1) Efficiency of parameter estimates. (2) Consistency of standard errors (var-cov matrix of β).

Two approaches: (1) GLS, generalized least squares. Model the clustering. What is the structure of Σ_m, the within-group correlation? (2) "Robust" formulation. Allow for arbitrary forms of clustering. Obtain consistent standard errors for any structure of Σ_m.

Example: Fixed Effects and Clusters

Example: The fixed effects panel data model

Decompose the error term $\nu_{mt} = u_m + e_{mt}$ so that the model is

$$y_{mt} = x_{mt}\beta + u_i + e_{mt}$$

and we assume the e_{mt} are *iid*. The structure of the within-group correlation is very special: since all the observations in a group share the same u_m, every observation within a group is equally-well correlated with every other observation. The structure of Σ_m in the block-diagonal var(ν) is σ_u^2 everywhere except the diagonal of the block, where it is $\sigma_u^2 + \sigma_e^2$.

The GLS approach is to use this model of the error structure. With the FE estimator, we partial-out the u_m by demeaning, and we're left with an *iid* idiosyncratic error e_{mt}. Note that this approach addresses both (1) efficiency of the estimate of β and (2) consistency of the estimated standard errors.

What's wrong with this approach? Nothing, if we've modeled the structure of the disturbances correctly. But if we haven't, then our estimated β isn't efficient – not such a problem – and our estimated standard errors are wrong – big problem!

The “Robust” Approach: Cluster-Robust Standard Errors

“Clustered errors” are an example of Eicker-Huber-White-“sandwich”-robust treatment of errors, i.e., make as few assumptions as possible. We keep the assumption of zero correlation across groups as with fixed effects, but allow the within-group correlation to be anything at all.

Some notation:

$$E(x_i' y_i) \equiv Q_{xy} \quad \hat{Q}_{xy} = \frac{1}{N} X'Y$$

$$E(x_i' x_i) \equiv Q_{xx} \quad \hat{Q}_{xx} = \frac{1}{N} X'X$$

Covariance matrix of orthogonality conditions (“GMM-speak”):

$$S = AVar(\overline{g}(\beta)) = \lim_{N\to\infty} \frac{1}{N} E(X'\nu\nu'X)$$

“Sandwich” variance matrix of β:

$$V = Q_{xx}^{-1} S Q_{xx}^{-1}$$

The “Robust” Approach: Cluster-Robust Standard Errors

“Sandwich” variance matrix of β:

$$V = Q_{xx}^{-1} S Q_{xx}^{-1}$$

Q_{xx} is estimated by $\hat{Q}_{xx}$. What will give $\hat{V}$ its robustness is our choice of the estimator $\hat{S}$.

If errors are *iid* (no robustness), then $S = \sigma^2 Q_{xx}$, we estimate $\hat{S}$ with $\hat{\sigma}^2 \hat{Q}_{xx}$ where $\hat{\sigma}^2$ is simply the root mean squared residual $\hat{\nu}$, and our estimate of the variance of β reduces to $\hat{V} = \hat{\sigma}^2 \hat{Q}_{xx}$, which is the standard, “classical” OLS variance estimator.

The “Robust” Approach: Cluster-Robust Standard Errors

“Sandwich” variance matrix of β:

$$V = Q_{xx}^{-1} S Q_{xx}^{-1}$$

If errors are independent but heteroskedastic, we use the Eicker-Huber-White-“robust” approach. $\hat{S} = \frac{1}{N}\sum_{i=1}^{N} x_i' x_i \hat{\nu}_i^2$ or, in matrix notation, $\hat{S} = \frac{1}{N} X' B X$ where B is a matrix with the squared residuals $\hat{\nu}_i^2$ running down the diagonal and zeros elsewhere. This estimate of S is robust to arbitrary heteroskedasticity, and therefore so is our estimate of V. The intuition is that although B (which looks like the covariance of ν) is NxN, S is KxK and fixed. We can't get a consistent estimate of the covariance of ν – you can't estimate an *nxn* matrix with only n observations – but we don't need it. We need only a consistent estimate of S, and with the number of observations N going off to infinity, the asymptotics give us this.

The “Robust” Approach: Cluster-Robust Standard Errors

The **cluster-robust** approach is a generalization of the Eicker-Huber-White-“robust” to the case of observations that are correlated within but not across groups. Instead of just summing across observations, we take the crossproducts of x and $\hat{\nu}$ for each group m to get what looks like (but isn't) a within-group correlation matrix, and sum these across all groups M:

$$\hat{S}_{CR} = \frac{1}{N} \sum_{m=1}^{M} \sum_{t=1}^{T_m} \sum_{s=1}^{T_m} x'_{mt} x_{ms} \hat{\nu}_{mt} \hat{\nu}_{ms}$$

The "Robust" Approach: Cluster-Robust Standard Errors

$$\hat{S}_{CR} = \frac{1}{N}\sum_{m=1}^{M}\sum_{t=1}^{T_m}\sum_{s=1}^{T_m} x'_{mt}x_{ms}\hat{\nu}_{mt}\hat{\nu}_{ms}$$

The intuition is similar to the heteroskedasticity-robust case. Since the within-group correlations are arbitrary and can vary from group to group, we can't estimate it with only one observation on each group. But we don't need this - we need only a consistent estimate of S, and if the number of groups M goes off to infinity, the asymptotics give us this.

This $\hat{S}_{CR}$ is consistent in the presence of **arbitrary within-group correlation** as well as arbitrary heteroskedasticity. This is what "cluster-robust" means in this context.

The “Robust” Approach: Cluster-Robust Standard Errors

Here’s an alternative exposition that highlights the parallels with the standard and heteroskedastic-robust covariance estimators.

General case: Covariance matrix of orthogonality conditions (“GMM-speak”):

$$S = AVar(\overline{g}(\beta)) = \lim_{N\to\infty} \frac{1}{N} E(X'\nu\nu' X)$$

Independently-distributed observations means $cov(\nu_i, \nu_j) = 0$, and S becomes $S = E(x_i' x_i \nu_i^2)$.

The “Robust” Approach: Cluster-Robust Standard Errors

Homoskedasticity means observations are identically as well as independently distributed (*iid*), and so $S = E(x_i' x_i \nu_i^2) = E(x_i' x_i) E(\nu_i^2)$. The standard estimator of S under the *iid* assumption is

$$\hat{S}_{homo} = \frac{1}{N} \sum_{i=1}^{N} x_i' x_i \quad \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}_i^2$$

where the second term is just the estimated error variance $\hat{\sigma}^2$ and the first term is just $\hat{Q}_{xx}$.

Heteroskedasticity means observations are not identically distributed, and we use instead the Eicker-Huber-White-robust estimator of S:

$$\hat{S}_{het} = \frac{1}{N} \sum_{i=1}^{N} x_i' x_i \hat{\nu}_i^2$$

The "Robust" Approach: Cluster-Robust Standard Errors

Now we consider "clustered" data. Observations are independent across clusters, but dependent within clusters. Denote by $X_{m.}$ the $T_m x K$ matrix of observations on X for the mth cluster, and $\nu_{m.}$ the $T_m x 1$ vector of disturbances for cluster m. Then we can write S as $S = E(X'_{m.}\nu_{m.}\nu'_{m.}X_{m.})$, where $E(\nu_{m.}\nu'_{m.})$ is just Σ_m, the covariance matrix of the disturbance ν for cluster m.

The cluster-robust covariance estimator for S is

$$\hat{S}_{CR} = \frac{1}{M}\sum_{m=1}^{M} X'_{m.}\hat{\nu}_{m.}\hat{\nu}'_{m.}X_{m.}$$

Note the parallels with $\hat{S}_{het}$. We are summing over clusters instead of individual observations; the Xs inside the summation are all the observations on a cluster instead of a single row of data; the term inside the Xs looks like (but isn't) the autocovariance of the disturbance for the cluster instead of what looks like (but isn't) the variance for the observation in the heteroskedastic-robust approach.

The “Robust” Approach: Cluster-Robust Standard Errors

In fact, there is another covariance estimator due to Kiefer (1980) that is robust to clustering but assumes homoskedasticity. To keep thing simple, assume that the dataset is a balanced panel, so that $T_m = T \quad \forall\ m$. If the data are homoskedastic, the TxT matrix $\Sigma_m = \Sigma \quad \forall\ m$ and we can estimate $\hat{\Sigma}$ by

$$\hat{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} \hat{\nu}_{m.} \hat{\nu}'_{m.}$$

The Kiefer covariance estimator is

$$\hat{S}_{Kiefer} = \frac{1}{M} \sum_{m=1}^{M} X'_{m.} \hat{\Sigma} X_{m.}$$

Again, note the parallels, this time with the usual homoskedastic estimator $\hat{S}_{homo}$. With $\hat{S}_{homo}$, we weight each observation with a scalar $\hat{\sigma}^2$, but since it's a scalar it can be pulled out of the summation; With $\hat{S}_{Kiefer}$, we weight each cluster with the matrix $\hat{\Sigma}$.

The "Robust" Approach: Cluster-Robust Standard Errors

What about efficiency of the OLS $\hat{\beta}$?

If the model is just-identified, OLS is still the efficient estimator. Why? We've assumed no structure at all for the intra-group correlations, and we have no extra information to bring to the estimate of S. With no additional assumptions or information, OLS is the best we can do.

If the model is overidentified, however, the cluster-robust approach can be used to obtain more efficient estimates of β via two-step or CUE (continuously-updated) GMM. This is the generalization of "heteroskedastic OLS" (Cragg 1983) to the case of clustered errors. "Overidentified" means that there are variables (instruments) that are not regressors, that are uncorrelated with ν, but that are "correlated" with the form of within-group clustering and/or heteroskedasticity in the data. The resulting $\hat{\beta}$ will be both consistent and efficient in the presence of arbitrary clustering and heteroskedasticity. In Stata, use `ivreg2` with the `cluster(id) gmm2s` options.

Example: Estimation of a System of Equations

We have a system of T equations. For each equation, we have M observations. Regressors are all exogenous. We want to be able to test cross-equation restrictions. "Clustering" arises because we use the same dataset to estimate all the equations, and the error ν_{mt} for the mth observation in equation t will be correlated with the error ν_{ms} for the mth observation in equation s.

The GLS approach is Zellner's "seemingly-unrelated regressions estimator" (*SURE*). We **model** the covariances of the ν_{mt}, estimate them, and construct the variance-covariance matrix that incorporates these off-diagonal elements. This lets us perform tests across equations, and obtain more efficient estimates of β in a second step. In Stata, this is the `sureg` command.

... but if we model the covariances incorrectly, all our inferences and testing will be wrong.

Example: Estimation of a System of Equations

The "robust" approach is to allow for **arbitrary** correlation of the ν_{mt} across equations. This uses the cluster-robust covariance estimator, where each observation m in the dataset defines a group or cluster. This is automated in Stata with the `suest` ("seemingly-unrelated estimations") command. It generates standard errors that are robust to heteroskedasticity as well as allowing cross-equation tests, but leaves the point estimates unchanged.

Alternatively, we can "stack" the equations "by hand" and use the cluster-robust covariance estimator. Estimation with OLS generates the same results as `suest`. However, if the model is overidentified (some regressors appear in one equation and not in others), we can do two-step GMM with `ivreg2` and obtain efficiency gains in our estimate of β.

Combining the GLS and Cluster-Robust Approaches

It is possible – and in some literatures, standard – to combine the GLS and cluster-robust approaches.

Consider again the fixed effects model. Say we consider it to be a good first approximation to within-group correlation, but there may be remaining within-group correlation even after accounting for fixed effects. For example, the e_{mt} could be serially correlated. One possibility would be to model the serial correlation, GLS-style. This is possible with the Stata command `xtregar`.

Alternatively, we could partial out the fixed effects in the usual way, and then use the cluster-robust covariance estimator. The only difference is that instead of using $\hat{\nu}_{mt}$ as residuals, we are using $\hat{e}_{mt}$. In effect, we are using cluster-robust to address any within-group correlation remaining after the fixed effects are removed. In Stata, use `xtreg,fe` with `cluster(id)`.

Combining the GLS and Cluster-Robust Approaches

First-differencing (FD) can be similarly motivated. With lagged dependent variables, we have to FD to get rid of the fixed effects (Arellano-Bond et al.). We then use cluster-robust errors to mop up the remaining and/or introduced serial correlation.

For some reason, combining the GLS and robust approaches is absolutely standard in the panel/serial correlation literature, and almost completely ignored in cross-section/heteroskedasticity practice. It's perfectly reasonable to do feasible GLS on a cross-section to get improvements in efficiency and then use robust SEs to address any remaining heteroskedasticity, but nobody seems to do this (GLS is too old-fashioned, perhaps).

Number of Clusters

The cluster-robust standard error estimator converges to the true standard error as the number of clusters M (not the number of observations N) approaches infinity.

Kézdi (2003) shows that 50 clusters (with roughly equal cluster sizes) is often close enough to infinity for accurate inference, and further that, even in the absence of clustering, there is little to no cost of using the CR estimator, as long as the number of clusters is large. 50 is an interesting number because of the many studies that use US state-level data.

With a small number of clusters ($M << 50$), or very unbalanced cluster sizes, the cure can be worse than the disease, i.e., inference using the cluster-robust estimator may be incorrect more often than when using the classical covariance estimator.

Rank of VCV

The rank of the variance-covariance matrix produced by the cluster-robust estimator has rank no greater than the number of clusters M, which means that at most M linear constraints can appear in a hypothesis test (so we can test for joint significance of at most M coefficients).

In a fixed-effect model, where there are a large number of parameters, this often means that test of overall model significance is feasible. However, testing fewer than M linear constraints is perfectly feasible in these models, though when fixed effects and clustering are specified at the same level, tests that involve the fixed effects themselves are inadvisable (the standard errors on fixed effects are likely to be substantially underestimated, though this will not affect the other variance estimates in general).

Estimates and their VCV

Note that the heteroskedasticity-robust and cluster-robust estimators for standard errors have no impact whatsoever on point estimates.

One **could** use information about the within-cluster correlation of errors to obtain more efficient estimates in many cases (see e.g. Diggle et al. 2002). There are also a variety of multi-level methods of parameterizing the distribution of errors to obtain more efficient estimates (using e.g. `xtmixed` and other model types—see Rabe-Hesketh and Skrondal 2005 for more). We will focus however on models where the **point estimates are unchanged** and only the estimated variance of our point estimates is affected by changing assumptions about errors.

In addition to improving the efficiency of the point estimates in regressions, modeling intra-cluster correlations can also result in improvements in meta-analysis, both in correctly modeling the variance of individual estimates and computing effect sizes. See Hedges (2006) for details.

Sandwich Estimators and Other Robustifications

Eicker (1967) and Huber (1967) introduced these sandwich estimators, but White (1980; 1982), Liang and Zeger (1986), Arellano (1987), Newey and West (1987), Froot (1989), Gail, Tan, and Piantodosi (1988), Kent (1982), Royall (1986), and Lin and Wei (1989), Rogers (1993), Williams (2000), and others explicated and extended aspects of the method in a non-survey context, so these are often cited as sources in specific applications. In the context of clustering induced by survey design, Kish and Frankel (1974), Fuller (1975), and Binder (1983), and Binder and Patak (1994), also derived results on cluster-robust estimators with broad applicability.

Stock and Watson (2006) point out that with fixed effects, both the standard heteroskedasticity-robust and HAC-robust covariance estimators are inconsistent for T fixed and $T > 2$, but the cluster-robust estimator does not suffer from this problem. One of their conclusions is that if serial correlation is expected, the cluster-robust estimator is the preferred choice.

Finite-Sample Adjustments

The cluster-robust covariance estimator is often used with a finite-sample adjustment q_c. The most common three forms are:

$$q_c = 1$$

$$q_c = \frac{N-1}{N-K}\frac{M}{M-1}$$

$$q_c = \frac{M}{M-1}$$

The Stata manual entry "Methods and Formulas" of [R] `regress` calls these the regression-like formula and the asymptotic-like formula, respectively. Fuller et al. (1986) and Mackinnon and White (1985) discuss finite-sample adjustments in more detail.

The Nature of the CR Correction

The heteroskedasticity-robust SE estimator scales not by the sum of squared residuals, but by the sum of "squared" products of residuals and the X variables, and the CR estimator further sums the products within cluster (if the products are negatively correlated within cluster, the CR standard errors will be smaller than the HR standard errors, and if positively correlated, larger). If the traditional OLS model is true, the residuals should, of course, be uncorrelated with the X variables, but this is rarely the case in practice.

The correlation may arise not from correlations in the residuals within a correctly specified model, but from specification error (such as omitted variables), so one should always be alert to that possibility.

Misspecification and the CR Correction

As Sribney (1998) points out: When CR estimates are smaller than standard SE estimates,

> *[S]ince what you are seeing is an effect due to (negative) correlation of residuals, it is important to make sure that the model is reasonably specified and that it includes suitable within-cluster predictors. With the right predictors, the correlation of residuals could disappear, and certainly this would be a better model.*
>
> *...[S]uppose that you measured the number of times each month that individuals took out the garbage, with the data clustered by household. There should be a strong negative correlation here. Adding a gender predictor to the model should reduce the residual correlations.*

The CR estimator will do nothing about bias in $\hat{\beta}$ when $E(X'e) \neq 0$.

Approximating the CR Correction

As Cameron, Gelbach, and Miller (2006a, p.5) note, if the primary source of clustering is due to group-level common shocks, a useful approximation is that for the kth regressor the default OLS variance estimate based on $s^2(X'X)^{-1}$ should be inflated by a factor of

$$1 + \rho_{\mathrm{e}}\rho_{\mathrm{x}_k}(\bar{N}_g - 1)$$

where ρ_{x_k} is the intra-cluster correlation of x_k, ρ_{e} is the intra-cluster correlation of residuals, and $\bar{N}_g$ is the average cluster size; in many settings the adjustment factor can be large even if ρ_{e} is small.

This approximation is closely related to the approximation given in Kish (1965, p.162) for the estimation of means in clustered data: he recommends inflating the variance estimate for the mean by a factor (or the SE by the square root of the factor):

$$1 + r(\bar{N}_g - 1)$$

where r is the measure of intraclass correlation (ICC) known as roh [not rho]. The approximation for regression with group-level common shocks is quite similar, with the adjustment that we now want the mean of y conditional on X.

Non-Nested and Nested Clusters

An extension of the basic one-dimensional case is to multiple levels of clustering. For example, errors may be clustered by country and by city, or errors may be clustered by country and by year. In the first case, the levels of clustering are nested, but in the second case, the clustering is along two dimensions and observations in each cluster along one dimension may appear in multiple clusters along the other. The latter case of non-nested clusters is discussed by Cameron, Gelbach, and Miller (2006a), who provide Stata code for estimating cluster-robust standard errors in this case.

To estimate cluster-robust standard errors in the presence of nested multi-level clustering, one can use the `svy` suite of commands.

Nested Clusters Using svy

It is straightforward to compute cluster-robust estimates for multi-level clustering with nested clusters using

```
svyset clevel1 || clevel2
```

(pweights are easily added as well) and then any command that allows the `svy:` prefix. In general, however, the correction at the highest level is the important one. Specifying clustering at the classroom level and clustering at the school level is unlikely to result in any substantive differences in inference relative to merely specifying clustering at the school level.

This argues for always specifying clustering at the highest of all nested levels at which intra-cluster correlation in errors may be a problem, but there is a tradeoff: at higher levels the number of clusters will be smaller, so the asymptotic results for the estimator are less likely to hold.

Problems with Cluster-Robust SEs

Why specify `cluster` (or use `svy`)?

- If the assumptions are satisfied, and errors are clustered, you'll get much better SE estimates.
- If the assumptions are satisfied, and errors aren't clustered, you'll get roughly the same SE estimates as if you had not specified `cluster` (i.e. no cost of robustness).

Why not always specify `cluster` (or use `svy`)?

- Convergence
- Bias
- Correlation across clusters
- Degrees of freedom

Speed of Convergence

The CR estimator is asymptotic in the number of clusters M. If M is small, there is no guarantee that the cluster-robust estimator will improve your inference—the cluster-robust estimator may make matters worse.

Kézdi (2003) shows that 50 clusters is often close enough to infinity for accurate inference, but these are simulations for a specific type of model. You may want to do simulations for a model that fits your specific application if you are worried about the convergence of the cluster-robust estimator, and what it implies for the reliability of your inferences.

Downward Bias

Rogers (1993) argues that “if no cluster is larger than 5 percent or so of the total sample, the standard errors will not be too far off because each term will be off by less than 1 in 400.” This implies that CR SEs with 20 equal-sized clusters would suffer from a very small bias.

With finite M, the cluster-robust estimator produces estimates of standard errors that are too small on average (i.e. they are biased downward). With M much less than 50, the bias can be substantial, particularly with $M < 10$. Cameron, Gelbach, and Miller (2006b) report that a “wild bootstrap” cluster-robust estimator performs well when $M < 50$. See also Wooldridge (2003) for more discussion and suggestions.

Degrees of freedom

Since the rank of the VCV matrix produced by the CR estimator is no greater than the number of clusters M you may not be able to test as many parameters as desired. For example, you could not cluster at the panel level and test for panel-specific intercepts and trends, since you would have at least twice as many parameters as degrees of freedom.

Given the limits on the number of parameters that may be tested in theory, even asymptotically, one might be worried about the small-sample properties of tests that involve nearly as many constraints as M. We will present simulations for certain cases.

A test for clustering

If you're worried about potential problems when using CR estimates, you'd like to test for the presence of clustering, to see whether you really need to adjust for clustering. Kézdi (2003) provides a test for clustering in the spirit of the White (1980) test for heteroskedasticity (see `hettest, whitetst, ivhettest` in Stata)

The intuition behind the Kézdi test is the same as that for the White test. The White general test for heteroskdesticity compares the $K(K+1)/2$ elements of the classical (unrobust) V with those of the heteroskedastic-robust V. If the difference is "large", the null of no heteroskedasticity is rejected. Similarly, the Kézdi test compares the elements of the classical V with the cluster-robust V. A "large" difference indicates the presence of clustering. Note that the tests reject when hypothesis testing involving β would be distorted, a very appealing property.

Balanced Panels, Equal Cluster Sizes, OLS-SE

Suppose we have the error components model

$$Y_{it} = X_{mt}\beta + u_m + e_{mt}$$

with $\beta_1 = 1$ and we have $M = 50$ balanced clusters, and $T = 20$ observations per cluster. Let the share of error variance due to the within-cluster component vary from 0 to 1 (across rows) and the share of within-cluster variation in regressors vary from 0 to 1 (across columns), and test $H_0 : \beta_1 = 1$ with $\alpha = 0.05$:

Rejection rates, nominal 5 percent level, OLS-SE

	0	25	50	75	100
0	.048	.043	.049	.048	.065625
25	.054	.057	.113	.157	.3052959
50	.052	.153	.312	.455	.6832814
75	.054	.209	.468	.679	.876161
100	.056	.241	.503	.716	

Balanced Panels, Equal Cluster Sizes, HRSE

Rejection rates, nominal 5 percent level, Het-Robust SE

	0	25	50	75	100
0	.049	.045	.05	.049	.0708333
25	.051	.057	.112	.154	.3094496
50	.054	.154	.321	.459	.6874351
75	.053	.202	.475	.679	.877193
100	.056	.242	.503	.715	

Balanced Panels, Equal Cluster Sizes, CR

Rejection rates, nominal 5 percent level, Clust-Robust SE

	0	25	50	75	100
0	.054	.039	.06	.09	
25	.053	.046	.107	.196	
50	.052	.07	.139	.335	
75	.056	.08	.179	.425	
100	.054	.078	.189	.434	

Balanced Panels, Equal Cluster Sizes, FECR

Rejection rates, nominal 5 percent level, FE and Clust-Robust SE

	0	25	50	75	100
0	.061	.038	.055	.055	
25	.054	.04	.044	.042	
50	.057	.054	.053	.062	
75	.056	.047	.044	.058	
100	.046	.047	.052	.042	

Unbalanced Panels and Unequal Cluster Sizes, OLS-SE

Now suppose we have 50 clusters and 1000 observations again, but 10 observations per cluster in 49 clusters and one cluster with 510 obs:

Rejection rates, nominal 5 percent level, OLS-SE

	0	25	50	75	100
0	.047	.056	.053	.058	.0679916
25	.047	.071	.073	.1	.1753112
50	.05	.171	.223	.347	.5658996
75	.04	.221	.41	.589	.8569948
100	.044	.27	.452	.677	

Unbalanced Panels and Unequal Cluster Sizes, HRSE

	0	25	50	75	100
0	.045	.053	.05	.059	.0700837
25	.048	.069	.077	.098	.1991701
50	.05	.166	.207	.34	.5774059
75	.047	.216	.388	.569	.8632124
100	.045	.271	.436	.654	

Unbalanced Panels and Unequal Cluster Sizes, CR

	0	25	50	75	100
0	.113	.104	.106	.123	
25	.105	.104	.095	.166	
50	.071	.133	.106	.253	
75	.031	.111	.096	.297	
100	.024	.116	.092	.299	

Unbalanced Panels and Unequal Cluster Sizes, FECR

	0	25	50	75	100
0	.119	.112	.115	.127	
25	.134	.123	.097	.111	
50	.106	.113	.103	.129	
75	.118	.118	.123	.126	
100	.088	.11	.078	.084	

Testing the Limits of df

Kézdi (2004) and our own simulations tell us that the CR estimator performs extremely well in relation to the HR or OLS SE estimators with respect to inference on a single parameter, as long as we have at least 50 clusters.

However, we know that we cannot test more than M coefficients. It makes sense to question how well the CR estimator performs when testing $M-2$ or $M-1$ coefficients.

Preliminary simulations show that the rejection rate rises from 5 percent to 100 percent as the number of coefficients increases from 1 to M. This needs further investigation.

Comparisons to a Parametric Correction

Suppose we have autocorrelated errors in a panel model:

$$Y_{it} = X_{it}\beta + u_i + e_{it}$$

with

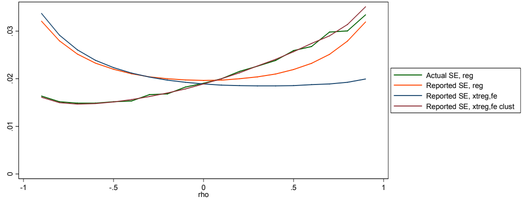
$$e_{it} = \rho e_{i(t-1)} + z_{it}$$

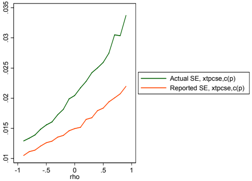
where z_{it} is *iid*. We could use `xtregar y x, fe`, `xtpcse y x, c(p)`, or `xtreg y x, fe cluster()`. How do these compare in finite samples? We can use MC simulation to evaluate the two approaches.

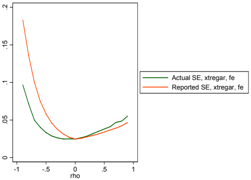
Additionally, Wooldridge (2002, pp.282-283) derives a simple test for autocorrelation in panel-data models, and the user-written program `xtserial` (Drukker 2003) performs this test in Stata. We can compare the performance of `xtserial` and `cltest` using MC simulation.

SE Estimates with Autocorrelation

Suppose $t \in \{1, 2, 3, 4, 5, 6, 7\}$ and $x = t - 4$ with $y = x + e$ and $M = 100$ (i.e. we are estimating a trend line $\beta = 1$ and there are 100 clusters). Suppose u_i is mean zero and uniform on $(-.5, .5)$. Here is a comparison of the reported and true SD of the OLS estimates (see also Diggle et al. 2002 Figure 1.7):





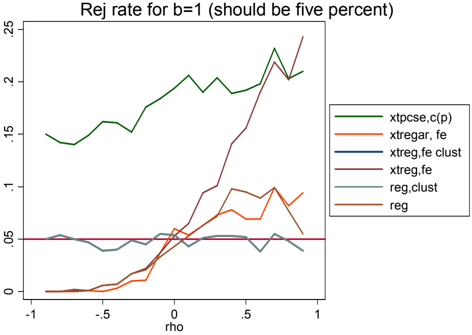


Rejection Rates, AR(1) Errors

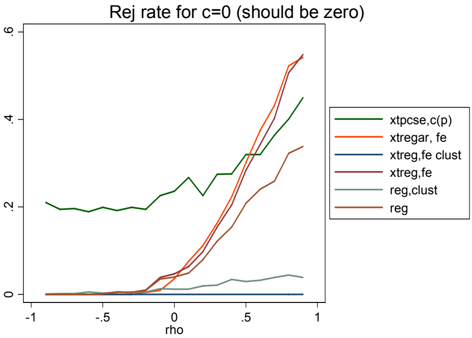
Mean rejection rates of $\beta = 1$ with nominal size 0.05

rho	reg	xtreg, fe	xtpcse	xtregar	reg, clust	xtreg, fe clust
-.9	0	0	.15	0	.05	.05
-.5	.006	.006	.162	0	.039	.039
-.1	.033	.037	.184	.037	.055	.055
0	.043	.053	.194	.06	.054	.054
.1	.053	.065	.206	.054	.043	.043
.5	.095	.156	.192	.069	.052	.052
.9	.055	.243	.21	.094	.039	.039

Rej Rates, AR(1) Errors

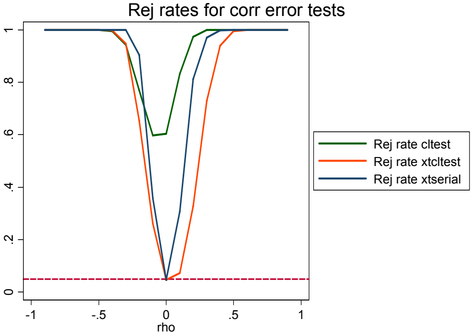


Rej Rates, AR(1) Errors



Tests for Clustering, AR(1) Errors

The test for clustering after `reg` is `cltest`, and the test for clustering after `xtreg, fe` is `xtcltest` (to be available from SSC shortly). It performs nearly as well as `xtserial` (which by construction is the correct test for this particular variety of clustering):



“Conclusions”

- Why not always use `cluster`?
- Small number of clusters
- Balanced vs unbalanced clusters
- Big clusters vs small clusters
- Number of hypotheses to test
- Testing for clustering (or heteroskedasticity)
- Efficiency gains by modeling the autocorrelation (GLS)
- Cluster AND GLS

More Examples and Simulations?

We plan to turn this talk into a Stata Journal submission. Any suggestions on additional topics that you feel should be included are welcomed— contact Austin at austinnichols@gmail.com or Mark at M.E.Schaffer@hw.ac.uk if you like.

References

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics,* 49: 431-34.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics,* 119(1): 249-275.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2007. "Enhanced routines for instrumental variables/GMM estimation and testing." Unpublished working paper, forthcoming.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2003. "Instrumental variables and GMM: Estimation and testing." *Stata Journal*, StataCorp LP, vol. 3(1), 1-31. Also Boston College Department of Economics Working Paper No 545

Binder, D. A. 1983. "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review,* 51: 279-292.

Binder, D. A. and Z. Patak. 1994. "Use of estimating functions for estimation from complex surveys." *Journal of the American Statistical Association,* 89(427): 10351043.

Cameron, Colin A., Jonah Gelbach, and Douglas L. Miller. 2006a. “Robust Inference with Multi-way Clustering.” NBER Technical Working Paper No. 327

Cameron, Colin A., Jonah Gelbach, and Douglas L. Miller. 2006b. “Bootstrap-Based Improvements for. Inference with Clustered Errors.”

Diggle, Peter J., Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger. 2002. *Analysis of Longitudinal Data, Second Edition.* Oxford University Press.

Drukker, David M. 2003. “Testing for serial correlation in linear panel-data models.” *Stata Journal,* (3)2: 1-10.

Eicker, F. 1967. “Limit theorems for regressions with unequal and dependent errors.” *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* Berkeley: University of California Press, 1: 59-82.

Froot, K. A. 1989. “Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data.” *Journal of Financial and Quantitative Analysis,* 24: 333355.

Fuller, W. A. 1975. “REgression analysis for sample survey.” *Sankhyā, Series C* 37: 117-132.

Gail, M. H., W. Y. Tan, and S. Piantodosi. 1988. “Tests for no treatment effect in randomized clinical trials.” *Biometrika* 75: 57-64.

Hedges, Larry V. 2006. "Effect Sizes in Cluster-Randomized Designs." IPR WP-06-13.

Huber, P. J. 1967. "The behavior of maximum likelihood estimates under non-standard conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley: University of California Press, 1, 221-233.

Kent, J. T. 1982. "Robust properties of likelihood ratio tests." *Biometrika,* 67: 19-27.

Kézdi, Gábor. 2004. "Robust Standard Error Estimation in Fixed-Effects Panel Models." *Hungarian Statistical Review* Special(9): 96-116.

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

Kish, Leslie and M. R. Frankel. 1974. "Inference from complex samples." *Journal of the Royal Statistical Society, Series B* 36: 1-37.

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

Lin, D. Y. and L. J. Wei. 1989. "The robust inference for the Cox proportional hazards model." *Journal of the American Statistical Association,* 84: 1074-1078.

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

Newey, W. K. and K. D. West. 1987. "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix." *Econometrica,* 55: 703708.

Rabe-Hesketh, Sophia and Anders Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata.* College Station, TX: Stata Press.

Rogers, William H. 1993. "sg17: Regression standard errors in clustered samples." *Stata Technical Bulletin* 13: 19-23.

Royall, R. M. 1986. "Model robust confidence intervals using maximum likelihood estimators." *International Statistical Review,* 54: 221-226.

Schaffer, Mark E. 2007. "cltest: Stata module to test for clustering of residuals."

Sribney, William. 1998. “Comparison of standard errors for robust, cluster, and standard estimators.” Stata website.

Stock, James H. and Mark W. Watson. 2006. “Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression.” NBER Technical Working Paper 323.

White, Halbert. 1980. “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.” *Econometrica,* 48: 817-830.

White, Halbert. 1982. “Maximum likelihood estimation of misspecified models.” *Econometrica,* 50: 1-25.

Williams, R. L. 2000. “A note on robust variance estimation for cluster-correlated data.” *Biometrics,* 56: 645646.

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

Wooldridge, J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press. Available from Stata.com’s bookstore.