Evaluating one-way and two-way cluster-robust covariance matrix estimates

Christopher F Baum[1] Austin Nichols[2] Mark E Schaffer[3]

[1]Boston College and DIW Berlin

[2]Urban Institute

[3]Heriot–Watt University, IZA and CEPR

German Stata Users Group Meeting, July 2011

The importance of cluster-robust standard errors

In working with linear regression models, researchers are increasingly likely to abandon the assumption of $i.i.d.$ errors in favor of a more realistic error structure. The use of 'robust' standard errors has become nearly ubiquitous in the applied literature.

There are many settings where allowing for heteroskedasticity at the level of the observation is warranted, but that single deviation from an $i.i.d.$ structure may not be sufficient to account for the behavior of the error process.

In the context of time series data and large-T asymptotics, one might naturally consider HAC standard errors: those robust to both heteroskedasticity and autocorrelation, familiar to economists as 'Newey–West' standard errors.

In this talk, we will consider how a broader set of assumptions on the error process may often be warranted, in the contexts of cross-sectional data of a hierarchical nature or in panel data.

The key concept to be considered is that of the *cluster-robust covariance matrix*, or *cluster VCE*, which relaxes the $i.i.d.$ assumption of independent errors, allowing for arbitrary correlation between errors within *clusters* of observations.

These clusters may represent some hierarchical relationship in a cross-section, such as firms grouped by industries, or households grouped by neighborhood. Alternatively, they may be the observations associated with each unit (or time period) in a panel dataset.

As discussed in prior talks by Nichols and Schaffer (UKSUG'07) and in recent work by Cameron and Miller (UC Davis WP, 2010), estimation of the VCE without controlling for clustering can lead to understated standard errors and overstated statistical significance. Just as the use of the classical (*i.i.d.*) VCE is well known to yield biased estimates of precision in the absence of the *i.i.d.* assumptions, ignoring potential error correlations within groups, or clusters, may lead to erroneous statistical inference.

The standard approach to clustering generalizes the 'White' (robust/sandwich) approach to a VCE estimator robust to arbitrary heteroskedasticity: in fact, `robust` standard errors in Stata correspond to cluster-robust standard errors computed from clusters of size one.

Simple one-way clustering In simple one-way clustering for a linear model, we consider that each observation ($i = 1, \ldots, N$) is a member of one non-overlapping cluster, g ($g = 1, \ldots, G$).

$$y_{ig} = \mathbf{x}'_{ig}\beta + u_{ig}$$

The error is assumed to be independent across clusters:

$$E(u_{ig}u_{jg'}\mathbf{x}_{ig}, \mathbf{x}_{jg'}) = 0$$

for $i \neq j$ unless $g = g'$.

How might this behavior of the error process arise?

Common shocks The simplest example of within-cluster correlation of errors arises when the errors are not *i.i.d.*, but rather contain a common time-invariant shock component as well as an idiosyncratic component:

$$u_{ig} = \nu_g + \zeta_{ig}$$

where ν_g is a common shock, or cluster-specific error, itself *i.i.d.*, and ζ_{ig} is an *i.i.d.* idiosyncratic error. This is equivalent to the error representation in the two-way error components model of panel data, but may just as well arise in a cross-sectional context.

As in random effects, $Var[u_{ig}] = \sigma_\nu^2 + \sigma_\zeta^2$ and $Cov[u_{ig}, u_{jg}] = \sigma_\nu^2, \ \forall i \neq j$. Here, *g* is indexing the panel unit and *i* is indexing time in the typical case where these are repeated observations on a single panel.

The *intraclass correlation*, common to all pairs of errors in a cluster, is

$$\rho_u = Corr[u_{ig}, u_{jg}] = \frac{\sigma_\nu^2}{(\sigma_\nu^2 + \sigma_\zeta^2)}$$

This constant within-cluster correlation is appropriate where observations within a cluster are *exchangeable*, in Stata parlance, with no implicit ordering. Individuals living in a household, families within a village or firms within an industry might follow this assumption.

If common shocks are the primary cause of error clustering, classical OLS standard errors are biased downward, and should be inflated by a factor taking the intraclass correlation into account. The inflation factor for a particular regressor's coefficient is also an increasing function of the within-cluster correlation of the regressor.

It is important to note that the bias in classical OLS standard errors is, in the general case, a function of both the intraclass correlation of the errors, ρ_u, and the intraclass correlation of the regressors, ρ_x. If either of the intraclass correlations is zero, OLS standard errors are OK. Conversely, the bias is worst when the intraclass correlations are high. Group-invariant regressors are a special case where this can be a big problem: for example, a dataset on individuals from different regions, where the local unemployment rate is a regressor.

In fact, if we had a dataset containing a number of clusters, regressors taking on constant values within those clusters, and errors that are entirely shared by cluster members, OLS estimation on these data is equivalent to estimating the model

$$\bar{y}_g = \mathbf{x}_g'\beta + \bar{u}_g$$

where $\bar{y}$ contains within-cluster averages of the dependent variable. There are really only G observations in the model, rather than N.

If OLS is applied to the individual data, for a constant regressor within-cluster, the true variance of an estimated coefficient is $(1 + \rho_u(N^* - 1))$ times larger than the classical OLS estimate, where ρ_u is the intraclass correlation and N^* is the number of observations in each cluster.

This problem has been known in the statistics literature since the 1960s (Kish); it was introduced in the economics literature by Moulton (*REStat*, 1990). Moulton demonstrated that in many settings this adjustment factor, and the consequent overstatement of precision, can be sizable even when ρ_u is fairly small. In his example, with $N = 18946$ and $G = 49$ (US states), $\hat{\rho}_u = 0.032$: a quite modest intrastate error correlation. With average group size of 387, the correction factor is 13.3, so that cluster-corrected standard errors are $\sqrt{13.3} = 3.7$ times larger for a state-level regressor than those computed by standard OLS techniques.

Panel/longitudinal data The preceding makes use of a simple specification where the within-cluster error ν_g is invariant within the cluster. But this is just an expositional simplification. Panel data error components estimators such as random or fixed effects or first-differences is unlikely to be a complete answer to the problem in many applications. With a within-panel time dimension (multiple observations on a household), or a within-panel spatial dimension (multiple counties in a state), the assumption of equi-correlated errors from the common shocks model is unlikely to be appropriate, as the strength of unit-specific autocorrelations will depend on their time difference or spatial difference.

For instance, in the case of $AR(1)$ errors $u_{it} = \rho u_{i,t-1} + \zeta_{it}$, the within-cluster error correlation becomes $\rho^{|t-\tau|}$ for observations dated t and τ, respectively. The decline in correlation for longer time spans implies that taking account of the presence of clustering will have a smaller effect than in the common shocks model.

The cluster-robust VCE estimator Cluster-robust VCE estimates are generalizations of the 'sandwich' method used to compute heteroskedasticity-robust standard errors (Stata's `robust` option), as developed by Eicker, Huber, White et al. The relationships between the different VCEs are perhaps easiest to see if we maintain the panel data notation of $i = 1...N$ to index panels and $t = 1...T$ to index observations within a panel, and assume a balanced panel structure. The cluster-robust estimate uses the sandwich estimator

$$VCE(\hat{\beta}) = (\mathrm{X}'\mathrm{X})^{-1}\hat{\Omega}(\mathrm{X}'\mathrm{X})^{-1}$$

where the filling of the sandwich is

$$\hat{\Omega} = \frac{1}{NT}\sum_{i=1}^{N}\mathrm{X}_i'\hat{u}_i\hat{u}_i'\mathrm{X}_i = \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t}^{T}\mathbf{x}_{it}\hat{u}_{it}\right)\left(\sum_{t}^{T}\mathbf{x}_{it}\hat{u}_{it}\right)'$$

with $\hat{u}_i \equiv (\hat{u}_{i1}\ldots\hat{u}_{iT})'$,and $\mathrm{X}_i \equiv (\mathbf{x}_{i1}\ldots\mathbf{x}_{iT})'$.

To see where this formula comes from, let's back up to the unclustered case, or, alternatively, one observation per cluster $i = 1...N$. As $t = 1$ for every observation, we can drop t and the total sample size is N.

Consider the population moments, $E(x_i u_i)$, where u_i are the error terms. The corresponding sample moments are

$$\bar{g}(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \widehat{g}_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \hat{u}_i$$

where $\hat{u}_i$ are the residuals computed from point estimates $\hat{\beta}$ and $\widehat{g}_i \equiv \mathbf{x}_i \hat{u}_i$.

The VCE of $\hat{\beta}$ is, by the sandwich formula,

$$V = Q_{xx}^{-1} \Omega Q_{xx}^{-1}$$

where Ω is the asymptotic variance of $\sqrt{N}\bar{g}$, i.e., i.e., $\sqrt{n}\bar{g}_n \rightarrow_d N(0, \Omega)$.

Estimating Q_{xx}^{-1} is easy: we simply use the sample analog $\hat{Q}_{xx}^{-1} \equiv (\frac{1}{N}\mathrm{X}'\mathrm{X})^{-1}$ The key question is, how do we estimate Ω? How we go about it depends directly on the assumptions we are willing to make.

Ω is the asymptotic variance of $\sqrt{N}\bar{g}$. Let's multiply out $\sqrt{N}\overline{g} * \sqrt{N}\overline{g}'$:

$$N\overline{g}\overline{g}' = \frac{1}{N}\sum_{i=1}^{N} g_i \sum_{i=1}^{N} g_i'$$

$$= \frac{1}{N}\left(x_1 u_1 ... + x_i u_i + ... + x_N u_N\right)\left(x_1 u_1 ... + x_i u_i + ... + x_N u_N\right)'$$

and after collecting terms:

$$N\overline{gg}' = \frac{1}{N}\sum_{i=1}^{N} x_i x_i' u_i^2 + 2\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N} x_i u_i x_j' u_j$$

If we make the assumption of independence, the double-sum term disappears in expectation. And if we assume conditional homoskedasticity as well, then

$$\Omega = E(g_i g_i') = E(x_i x_i' u_i^2) = E(x_i x_i')E(u_i^2) = Q_{xx}\sigma_u^2$$

and the natural estimator is the usual classical estimator of Ω:

$$\hat{\Omega} = \hat{Q}_{xx} s^2$$

where s^2 is a consistent estimate of σ_u^2. Plugging this into the sandwich formula $\hat{V} = \hat{Q}_{xx}^{-1}\hat{\Omega}\hat{Q}_{xx}^{-1}$ gives us our $\widehat{V}$.

Return to $\sqrt{N}\overline{g} * \sqrt{N}\overline{g}'$:

$$N\overline{g}\overline{g}' = \frac{1}{N}\sum_{i=1}^{N} x_i x_i' u_i^2 + 2\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N} x_i u_i x_j' u_j$$

Maintain the assumption of independence (so the double-sum still disappears in expectation) relax the assumption of conditional homoskedasticity. The natural estimator of Ω for this set of assumptions is the Eicker–Huber–White 'robust' estimator

$$\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \hat{u}_i^2$$

with the VCE estimator as

$$\hat{V} = \hat{Q}_{xx}^{-1}\hat{\Omega}\hat{Q}_{xx}^{-1}$$

This is the VCE invoked by the `robust` option in Stata.

We can now finally return to the cluster-robust estimator.

Consider again Ω, the asymptotic variance of $\sqrt{N}\bar{g}$, but bring back in the clustered structure of the data, so that there is more than one observation per cluster and the time subscripts reappear:

$$
\begin{aligned}
N T \overline{gg}' &= \frac{1}{NT}\left(x_{11}u_{11}\ldots + x_{1T}u_{1T}x_{1T}u_{1T} + \ldots + x_{it}u_{it} + \ldots + x_{NT}u_{NT}\right) \\
&\quad \times \left(x_{11}u_{11}\ldots + x_{1T}u_{1T}x_{1T}u_{1T} + \ldots + x_{it}u_{it} + \ldots + x_{NT}u_{NT}\right)' \\
&= \frac{1}{NT}\left(\sum_{t=1}^{T} x_{1t}u_{1t} + \ldots + \sum_{t=1}^{T} x_{it}u_{it} + \ldots + \sum_{t=1}^{T} x_{Nt}u_{Nt}\right) \\
&\quad \times \left(\sum_{t=1}^{T} x_{1t}u_{1t} + \ldots + \sum_{t=1}^{T} x_{it}u_{it} + \ldots + \sum_{t=1}^{T} x_{Nt}u_{Nt}\right)' \\
&= \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t=1}^{T} x_{it}u_{it}\right)\left(\sum_{t=1}^{T} x_{it}u_{it}\right)' \\
&\quad + 2\frac{1}{NT}\sum_{i}^{N}\sum_{j\neq i}^{N}\left(\sum_{t=1}^{T} x_{it}u_{it}\right)\left(\sum_{t=1}^{T} x_{jt}u_{jt}\right)'
\end{aligned}
$$

The first set of terms is all the within-cluster cross-products. The second set of terms is all the between-cluster cross-products. In expectations, everything in the second set of terms disappears because all the between-cluster correlations are zero by assumption, leaving us with just the first set of terms.

This is the motivation for the cluster-robust covariance estimator of Ω:

$$\hat{\Omega} = \frac{1}{NT} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} x_{it} \widehat{u}_{it} \right) \left(\sum_{t=1}^{T} x_{it} \widehat{u}_{it} \right)'$$

Note that we can interpret the clusters as 'super-observations'. Furthermore, in the special case where errors are heteroskedastic but still independently distributed, the number of clusters N is equal to the number of observations, NT, and each cluster has one observation. In this case the cluster-robust formula reduces to the standard heteroskedasticity-robust Eicker–Huber–White formula implemented by Stata's `robust` option, as presented above. Finally, recall that the asymptotics here are holding T fixed and sending $N \to \infty$.

Bias in the cluster-robust estimator

While the formula for $\hat{\Omega}$ is appropriate as the number of clusters N goes to infinity, finite-sample corrections are usually applied to deal with downward bias in the cluster-robust standard errors. Stata uses $\sqrt{c}\hat{u}_i$ in computing $\hat{\Omega}$, with $c \simeq \frac{N}{N-1}$. Simulations have shown that the bias is larger when clusters are unbalanced: for instance, in a dataset with 50 clusters, in which half the data are in a single cluster and the other 49 contain about one percent of the data. A further finite-sample adjustment factor $\frac{NT-1}{NT-K}$ can also be applied.

As a rule of thumb, Nichols and Schaffer (2007) suggest that the data should have at least 20 balanced clusters or 50 reasonably balanced clusters. Rogers' seminal work (*Stata Tech.Bull.*, 1993) suggested that no cluster should contain more than five per cent of the data.

Cluster-robust t and F tests When a cluster-robust VCE has been calculated, Wald t or F test statistics should take account of the number of clusters, rather than relying on the asymptotically behavior of the statistic as $NT \rightarrow \infty$. The approach that Stata follows involves using the t distribution with $N - 1$ degrees of freedom rather than $NT - k$ degrees of freedom. If the number of clusters is small, this will substantially increase the critical values relative to those computed from the standard Normal (t with large d.f.).

Some authors (e.g., Donald and Lang (*Rev.Ec.Stat.*, 2007) recommend using t_{N-L}, where L is the number of regressors constant within cluster, as an even more conservative approach.

By what shall we cluster? In many microeconometric datasets there may be several choices for clustering. In cross-sectional individual-level data, we may consider clustering at the household level, assuming that individuals' errors will be correlated with those of other household members, but may also cluster at a higher level of aggregation such as neighborhood, city or state. With nested levels of clustering, clusters should be chosen at the most aggregate level (e.g., at the state level) to allow for correlations among individuals at that level. This advice must be tempered with the concern that a reasonable number of clusters is defined.

Moving away from pure cross-sectional data to the realm of pooled cross-section time-series data, we should consider alternative assumptions on the independence of errors over the time dimension.

For instance, individuals' errors may be clustered at the level of household, city or state, but clustering on one of those variables assumes that a common intraclass correlation applies to all pairs of errors belonging to individuals in the cluster over time. As discussed earlier, this may make sense in the unit dimension, but is less sensible in the time dimension.

Conversely, clustering may be defined for a given aggregation and time period: e.g., in a household study, at the state-year level. However, this form of clustering maintains the assumption that for a given state, individuals' errors are independent over time. This may be quite unrealistic, given the existence of state-level variables that have sizable correlations over time, even if they exhibit variation at the individual level (such as marginal tax rates).

This issue would be similarly relevant if we worked with firm-level panel data where clustering was defined at the industry-year level. High autocorrelations among industry-level measures would tend to invalidate the assumptions that errors for an industry are uncorrelated over time. If the clustering scheme was defined only in terms of industry, no restrictions would be placed on those correlations.

In panel data where we cluster by the unit identifier (e.g., firm id code), we allow for within-firm error correlations, but rule out across-firm error correlations such as those arising from common shocks. On the other hand, clustering by time period allows for common shocks, but assumes that errors associated with a given firm are independently distributed: a questionable assumption. One-way clustering by either firm or time period has its limitations.

In some cases, one-way clustering may be adequate: with errors clustered by firms and by year, the latter error correlations might be completely due to common shocks. In that case, the introduction of time fixed effects would absorb all within-year clustering, and one-way clustering on firms would be appropriate. However, if these shocks have a meaningful firm-level component, contemporaneous error correlations across firms will remain.

These concerns naturally lead to the generalization of the cluster-robust estimator to two or more dimensions.

Two-way clustering One-way clustering relies on the assumption that $E(u_i u_j \mathbf{x}_i \mathbf{x}_j') = 0$ unless observations i, j belong to the same cluster. In two-way clustering, the same assumption is made, and the matrix $\hat{\Omega}$ defined earlier is generalized to

$$\hat{\Omega} = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathrm{I}(it, js) \left[\mathbf{x}_{it} \mathbf{x}_{js}' \; \hat{u}_{it} \hat{u}_{js} \right]$$

where $\mathrm{I}(it, js) = 1$ for observations in the same cluster, and 0 otherwise.

Computation of the two-way cluster-robust VCE is straightforward, as Thompson (2010) illustrates. The VCE may be calculated from

$$VCE(\hat{\beta}) = VCE_1(\hat{\beta}) + VCE_2(\hat{\beta}) - VCE_{12}(\hat{\beta})$$

where the three VCE estimates are derived from one-way clustering on the first dimension, the second dimension and their intersection, respectively. As these one-way cluster-robust VCE estimates are available from most Stata estimation commands, computing the two-way cluster-robust VCE involves only a few matrix manipulations.

This procedure has been automated in Baum, Schaffer, Stillman's `ivreg2` and Schaffer's `xtivreg2` routine on SSC, which may be employed to estimate OLS models as well as models employing instrumental variables, IV-GMM and LIML.

One concern that arises with two-way (and multi-way) clustering is the number of clusters in each dimension. With one-way clustering, we should be concerned if the number of clusters N is too small to produce unbiased estimates. The theory underlying two-way clustering relies on asymptotics in *both* dimensions, i.e., both N and T. The two-way clustering approach is thus sensible only if there is a sizable number of clusters in each dimension.

Just as in one-way clustering, finite-sample adjustments should be made for the number of clusters. One approach, followed by Cameron et al.'s `cgmreg` routine, adjusts each of the three covariance matrices by a ratio reflecting the number of clusters in that matrix.

An alternate approach, implemented in `ivreg2`, computes $VCE(\hat{\beta})$ and then scales by $\frac{M}{M-1}$, where $M = \min(G_1, G_2)$ and G_1 and G_2 are the number of clusters in the two dimensions. Both approaches can also include a finite-sample adjustment factor based on the number of regressors K. In `ivreg2`, both adjustment factors are invoked with the `small` option.

We must keep in mind that the cluster-robust concept is much more general than the panel data setting. For instance, we may have firm-level data, categorized by both industry and region, and we may doubt the independence of errors within industry (for firms in different regions) as well as within region (for firms in different industries).

If we created a single clustering variable from the intersection of industries and regions, we would allow for error correlations between firms that were both in industry *i* and region *j*, and rule out correlations among all other pairs of firms: possibly an overly restrictive approach.

Revisiting the two-way clustering formula, you can see that one-way clustering by the intersection of the two dimensions would correspond to the third term in the formula, $VCE_{12}(\hat{\beta})$, whereas full two-way clustering by industry and region would allow for correlated errors across those dimensions as well.

Note, however, that we have to add an independence assumption (Kolenikov–Nichols 2011). Not only do we have to believe in independence across clusters in the first dimension (cluster *i* vs. cluster *j*), and in the second dimension (cluster *t* vs. cluster *s*), we also have to believe in independence across the two dimensions combined (observation *it* vs. observation *js*). Kolenikov and Nichols (2011) also demonstrate that in practice, the two-way clustering VCE may not be positive definite, even with large numbers of observations in both dimensions.

Multi-way clustering With that caveat in mind, we may extend the notion of cluster-robust VCEs to three or more non-nested dimensions. Multi-way clustering is described by Cameron, Gelbach, Miller [CGM] (*JBES*, 2011; UC Davis WP 09-9). For instance, we might consider data on individual workers, clustered by industry, occupation and US state.

The logic to compute the $\hat{\Omega}$ matrix, as CGM show, is a generalization of the formula for two-way clustering, and may be implemented using only one-way cluster-robust estimates available from many Stata estimation commands. Alternatively, CGM provide the `cgmreg` command, downloadable from
`http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/`
which implements multi-way clustering for linear regressions.

But because the asymptotic requirements increase with each dimension (each dimension is $\rightarrow \infty$), practical applications of multi-way clustering will be limited.

HAC vs. cluster-robust methods In the pure time-series context, the HAC ('kernel-robust', e.g. Newey–West/Bartlett kernel) estimator of the VCE allows for arbitrary serial correlation. The HAC estimator requires large-T asymptotics. The HAC VCE estimator can be easily applied in the panel data context; it amounts to applying the kernel-robust approach to each panel. Note that the arbitrary serial correlation is addressed using large-T asymptotics, and we need a panel with a long time series for this to work. The cross-section dimension can be small.

Contrast this with the simple one-way cluster-robust VCE, where we cluster in the N-dimension, on panel units. In that case, we obtain a VCE that is robust to arbitrary serial correlation using large-N asymptotics, and we need a dataset with a lot of observations in the cross-section dimension for this to work. The time-series dimension can be small.

Note that in both approaches, we are still assuming independence across panel units.

Can these approaches be combined? Yes, if we cluster in the T-dimension, on time periods If we employ clustering by time period, we can relax the independence assumption and allow for common shocks across panel units. Combining clustering on time with the HAC kernel-robust method means that we can obtain a VCE that allows for arbitrary shocks that are shared by panel units, and which arbitrarilyy serially correlated. Note that this relies on large-T asymptotics.

We can go further (Thompson, 2010) and combine two-way clustering and HAC. This provides SEs and statistics that are robust to autocorrelated within-panel disturbances (clustering on panel id) and to autocorrelated across-panel disturbances (clustering on time combined with kernel-based HAC). Now, however, we also require large-N asymptotics.

Heteroskedasticity and cluster-robust vs. cluster-only-robust The cluster-robust VCE discussed so far is robust to both clustering and heteroskedasticity. In fact, a VCE that is robust to clustering only and that assumed homoskedasticity is available. This is Kiefer's (1980) VCE :

$$\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} x_i \widehat{V}_u x_i'$$

$$\widehat{V}_u = \frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i \widehat{u}_i'$$

where $\widehat{V}_u$ is an estimate of the $T \times T$ covariance matrix of u_i, and u_i is the column of T errors for observation i.

The Kiefer VCE has not been widely used, partly for the same reasons that HAC covariance estimators are much more widely used than AC estimators that assume homoskedasticity, and partly because the generalization to unbalanced panels is tricky. But the Kiefer cluster-only-robust VCE is useful in a specification testing setting, a point made by Kézdi and to which we return below.

Fixed effects models with clustering In any context where we identify clusters, we could consider including a fixed-effect parameter for each cluster, as in

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}$$

As is well known from analysis of this model in the special case of longitudinal or panel data, the inclusion of the α_i parameters centers each cluster's residuals around zero. However, the inclusion of these fixed-effect parameters is a solution the intra-cluster correlation of errors only in the very special case of time-invariant correlation. For this reason, it is usually advisable to question the *i.i.d.* error assumption and produce cluster-robust estimates of the VCE.

Another reason to use the cluster-robust VCE is because it is robust to heteroskedasticity as well as autocorrelation. In fact, Stock–Watson (2008) have shown that the naive Eicker–Huber–White heteroskedasticity-robust VCE is *not* a consistent estimate of V for $T > 2$. (The reason is the same incidental parameters problem that makes the naive classical VCE inconsistent, and which is addressed by a dof adjustment.) For this reason, Stata's official `xtreg` automatically reports the cluster-robust VCE when the user specifies the fixed effects estimator combined with the `robust` option.

Note, however, that the Stock–Watson results also show that the bias in the naive Eicker–Huber–White heteroskedasticity-robust VCE.for the fixed-effects estimator disappears as T gets large. Thus a user with a large-T small-N panel with fixed effects may prefer this VCE estimator to the cluster-robust one. Stock and Watson also provide a bias-corrected heteroskedasticity-robust VCE.for the fixed effects estimator. Both are available in Schaffer's `xtivreg2` routine, the latter via the undocumented `sw` option.

Testing for cluster effects We might naturally wish to test whether the computation of the cluster-robust VCE is warranted, as in the case of 'robust' standard errors, the classical VCE estimate is to be preferred if *i.i.d.* assumptions are satisfied.

For the case of one-way clustering in fixed-effects panel models, Kézdi (*Hungarian Stat. Rev.*, 2004) presents a test based on White's (*Econometrica*, 1980) direct test for heteroskedasticity. The motivation of the original White test for heteroskedasticity is that, under the null of conditional homoskedasticity, the difference between the classical and heteroskedasticity-robust $\widehat{\Omega}$ should disappear as the sample gets large:

$$\widehat{\Omega}_{HC} - \widehat{\Omega}_{classical} = \left[\frac{1}{N} \sum_{i=1}^{N} x_i x_i' \widehat{u}_i^2 - \widehat{Q}_{xx} s^2 \right] \rightarrow_p 0$$

A quadratic form in the vector of contrasts yields a test statistic distributed χ^2 under the null of conditional homoskedasticity; a large test statistic indicates rejection of the null in favor of the alternative of heteroskedasticity.

Kézdi's test uses the same reasoning. Underlying the test is the point that, under the null of conditional homoskedasticity and independence, the difference between the classical and cluster-robust $\widehat{\Omega}$ should disappear as the sample gets large:

$$\widehat{\Omega}_{CR} - \widehat{\Omega}_{classical} = \frac{1}{NT} \sum_{i=1}^{N} \left[\left(\sum_{t=1}^{T} x_{it} \widehat{u}_{it} \right) \left(\sum_{t=1}^{T} x_{it} \widehat{u}_{it} \right)' - s^2 \sum_{t=1}^{T} x_{it} x_{it}' \right] \rightarrow_p 0$$

Kézdi's specific application was to the fixed effects estimator, but it can equally easily be applied to other settings. His study of his test's properties suggests that it performs well, even in the common 'small T, large N' setting, and also is reliable in models where T becomes large.

Kézdi's test can be easily extended to other specification tests by choosing the $\widehat{\Omega}$s appropriately:

- heteroskedasticity+clustering vs. homoskedasticity+independence ($\widehat{\Omega}_{CR} - \widehat{\Omega}_{classical}$)
- heteroskedasticity+clustering vs. heteroskedasticity only ($\widehat{\Omega}_{CR} - \widehat{\Omega}_{HR}$)
- clustering vs. homoskedasticity+independence ($\widehat{\Omega}_{Kiefer} - \widehat{\Omega}_{classical}$)
- heteroskedasticity+clustering vs. clustering-only ($\widehat{\Omega}_{CR} - \widehat{\Omega}_{Kiefer}$)

We have implemented a preliminary version of the Kézdi test for the hypothesis that the errors are *i.i.d.* versus the alternative that they exhibit within-cluster dependence as Stata command `chatest`, with the panel counterpart `xtchatest`.

A note in passing: White's general test, and Kézdi's cluster version, are based on a vector of contrasts constructed from *every* element of the VCE. Note that, in effect, we are looking for evidence of heteroskedasticity or clustering in every possible direction that could bias our VCE. But researchers are rarely equally interested in all their parameters and all possible combinations of tests of these parameters; usually they are interested in just one or two. Heteroskedasticity or clustering that affects inference involving the parameters of interest is a worry; if it affects inference involving the other parameters, so what? (More about this below, time permitting.)

Some empirical examples

Compare VCE estimates from a cross-section dataset computed under assumptions:

- *i.i.d.*
- robust
- cluster-robust by industry (9 categories)
- cluster-robust by occupation (9 categories)
- two-way cluster-robust

Table: Wage equation using modified nlsw88

	(1) iid	(2) robust	(3) clus_ind	(4) clus_occ	(5) clus_2way
hours	0.0545***	0.0545***	0.0545**	0.0545**	0.0545**
	(0.0114)	(0.0113)	(0.0166)	(0.0222)	(0.0199)
ttl_exp	0.268***	0.268***	0.268***	0.268***	0.268***
	(0.0260)	(0.0250)	(0.0387)	(0.0439)	(0.0471)
black	-0.696**	-0.696***	-0.696**	-0.696*	-0.696*
	(0.272)	(0.251)	(0.301)	(0.330)	(0.317)
collgrad	3.170***	3.170***	3.170***	3.170***	3.170***
	(0.274)	(0.314)	(0.443)	(0.643)	(0.491)
south	-1.365***	-1.365***	-1.365***	-1.365***	-1.365***
	(0.243)	(0.239)	(0.267)	(0.370)	(0.318)
N	2141	2141	2141	2141	2141

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

```
. ivreg2 wage hours ttl_exp black collgrad south, cluster(industry)
OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on industry
Number of clusters (industry) =     12                  Number of obs =     2228
                                                        F(  5,    11) =    58.79
                                                        Prob > F      =   0.0000
Total (centered) SS     =  74036.56905                  Centered R2   =   0.1537
Total (uncentered) SS   =    209279.49                  Uncentered R2 =   0.7006
Residual SS             =    62656.361                  Root MSE      =    5.303
------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0553955   .0145375     3.81   0.000     .0269025    .0838885
     ttl_exp |   .2704925   .0363919     7.43   0.000     .1991657    .3418193
       black |  -.7172461   .2771599    -2.59   0.010     -1.26047   -.1740226
    collgrad |   3.112953   .3704258     8.40   0.000     2.386932    3.838974
       south |  -1.370694   .2451673    -5.59   0.000    -1.851213   -.8901752
       _cons |    2.35104   .6115602     3.84   0.000     1.152404    3.549676
------------------------------------------------------------------------------
Included instruments: hours ttl_exp black collgrad south
------------------------------------------------------------------------------

. chatest, cluster(industry)
Test of:
heteroskedasticity (nR2, homokurt assumed)=        30.483  Chi-sq(20) p=0.0624
heteroskedasticity (no homokurt assumed)=          62.893  Chi-sq(20) p=0.0000
heteroskedasticity & clustering=                20465.921  Chi-sq(20) p=0.0000
heteroskedasticity & clustering (no centering)=    12.000  Chi-sq(20) p=0.9161
het+clustering vs het-only (centering)=          5958.146  Chi-sq(20) p=0.0000
het+clustering vs het-only (no centering)=         12.000  Chi-sq(20) p=0.9161
```

```
. ivreg2 wage hours ttl_exp black collgrad south, cluster(occupation)
OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on occupation
Number of clusters (occupation) =     13              Number of obs =     2233
                                                      F(  5,    12) =    10.83
                                                      Prob > F      =   0.0004
Total (centered) SS     =  74139.62742                Centered R2   =   0.1549
Total (uncentered) SS   =   209492.595                Uncentered R2 =   0.7009
Residual SS             =  62658.35322                Root MSE      =    5.297
------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0554197   .0201462     2.75   0.006     .0159339    .0949056
     ttl_exp |   .2696493   .0417182     6.46   0.000     .1878831    .3514154
       black |  -.7175845   .2874904    -2.50   0.013    -1.281055   -.1541137
    collgrad |   3.129628    .574179     5.45   0.000     2.004258    4.254999
       south |   -1.39033   .3313849    -4.20   0.000    -2.039832    -.740827
       _cons |   2.360572   .6004096     3.93   0.000     1.183791    3.537353
------------------------------------------------------------------------------
Included instruments: hours ttl_exp black collgrad south
------------------------------------------------------------------------------

. chatest, cluster(occupation)
Test of:
heteroskedasticity (nR2, homokurt assumed)=        30.513  Chi-sq(20) p=0.0620
heteroskedasticity (no homokurt assumed)=          63.035  Chi-sq(20) p=0.0000
heteroskedasticity & clustering=               261161.030  Chi-sq(20) p=0.0000
heteroskedasticity & clustering (no centering)=    13.000  Chi-sq(20) p=0.8774
het+clustering vs het-only (centering)=        260239.118  Chi-sq(20) p=0.0000
het+clustering vs het-only (no centering)=         12.000  Chi-sq(20) p=0.9161
```

```
. qui reg wage ttl_exp black collgrad south
. predict double wage_e, resid
. qui reg hours ttl_exp black collgrad south
. predict double hours_e, resid
(4 missing values generated)
. ivreg2 wage_e hours_e, cluster(industry) nocons
OLS estimation
------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on industry
Number of clusters (industry) =     12                  Number of obs =      2228
                                                        F(  1,    11) =     13.34
                                                        Prob > F      =    0.0038
Total (centered) SS     =   63362.8718                  Centered R2   =    0.0111
Total (uncentered) SS   =  63363.24095                  Uncentered R2 =    0.0111
Residual SS             =  62657.02256                  Root MSE      =     5.303
------------------------------------------------------------------------------
             |               Robust
      wage_e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hours_e |   .0553973   .0145242     3.81   0.000     .0269304    .0838642
------------------------------------------------------------------------------
Included instruments: hours_e
------------------------------------------------------------------------------

. chatest, cluster(industry)
Test of:
heteroskedasticity (nR2, homokurt assumed)=          0.005  Chi-sq( 1) p=0.9430
heteroskedasticity (no homokurt assumed)=            0.013  Chi-sq( 1) p=0.9087
heteroskedasticity & clustering=                     1.132  Chi-sq( 1) p=0.2873
heteroskedasticity & clustering (no centering)=      0.903  Chi-sq( 1) p=0.3419
het+clustering vs het-only (centering)=              1.430  Chi-sq( 1) p=0.2318
het+clustering vs het-only (no centering)=           1.050  Chi-sq( 1) p=0.3055
```

```
. ivreg2 wage_e hours_e, cluster(occupation) nocons
OLS estimation
------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on occupation
Number of clusters (occupation) =     13                  Number of obs =      2233
                                                          F(  1,    12) =      6.95
                                                          Prob > F      =    0.0217
Total (centered) SS     =  63366.41498                    Centered R2   =    0.0112
Total (uncentered) SS   =  63366.59346                    Uncentered R2 =    0.0112
Residual SS             =  62659.06348                    Root MSE      =     5.297
------------------------------------------------------------------------------
             |               Robust
      wage_e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hours_e |   .0554236   .0201922     2.74   0.006     .0158476    .0949995
------------------------------------------------------------------------------
Included instruments: hours_e
------------------------------------------------------------------------------

. chatest, cluster(occupation)
Test of:
heteroskedasticity (nR2, homokurt assumed)=          0.004  Chi-sq( 1) p=0.9491
heteroskedasticity (no homokurt assumed)=            0.010  Chi-sq( 1) p=0.9185
heteroskedasticity & clustering=                     3.173  Chi-sq( 1) p=0.0749
heteroskedasticity & clustering (no centering)=      2.448  Chi-sq( 1) p=0.1177
het+clustering vs het-only (centering)=              3.779  Chi-sq( 1) p=0.0519
het+clustering vs het-only (no centering)=           2.869  Chi-sq( 1) p=0.0903
```

Some empirical examples

Compare VCE estimates from a panel dataset computed under assumptions:

- *i.i.d.*
- cluster-robust by company (10 units)
- two-way cluster-robust (10 companies and 20 time periods)

Table: Investment equation using grunfeld

	(1) iid	(2) clus_comp	(3) clus_2way
mvalue	0.110***	0.110***	0.110***
	(0.0119)	(0.0152)	(0.0117)
kstock	0.310***	0.310***	0.310***
	(0.0174)	(0.0528)	(0.0435)
N	200	200	200

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

```
. ivreg2 invest mvalue kstock
OLS estimation
------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =       200
                                                      F(  2,   197) =    426.58
                                                      Prob > F      =    0.0000
Total (centered) SS     =  9359943.917                Centered R2   =    0.8124
Total (uncentered) SS   =  13620706.07                Uncentered R2 =    0.8711
Residual SS             =  1755850.432                Root MSE      =      93.7

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1155622   .0057918    19.95   0.000     .1042105    .1269138
      kstock |   .2306785    .025284     9.12   0.000     .1811227    .2802342
       _cons |  -42.71437   9.440069    -4.52   0.000    -61.21656   -24.21217
------------------------------------------------------------------------------
Included instruments: mvalue kstock
------------------------------------------------------------------------------

. chatest, cluster(company)
Test of:
heteroskedasticity (nR2, homokurt assumed)=        91.790  Chi-sq( 5) p=0.0000
heteroskedasticity (no homokurt assumed)=          33.377  Chi-sq( 5) p=0.0000
heteroskedasticity & clustering=                  406.674  Chi-sq( 5) p=0.0000
heteroskedasticity & clustering (no centering)=     4.459  Chi-sq( 5) p=0.4854
het+clustering vs het-only (centering)=            39.595  Chi-sq( 5) p=0.0000
het+clustering vs het-only (no centering)=          8.214  Chi-sq( 5) p=0.1448
clustering vs homosked (no centering)=              8.855  Chi-sq( 5) p=0.1150
het+clustering vs clust-only=                      10.684  Chi-sq( 5) p=0.0580
het+clustering vs clust-only (no centering)=        9.646  Chi-sq( 5) p=0.0859
```

```
. ivreg2 invest mvalue kstock
OLS estimation
------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      200
                                                      F(  2,   197) =   426.58
                                                      Prob > F      =   0.0000
Total (centered) SS     =  9359943.917                Centered R2   =   0.8124
Total (uncentered) SS   =  13620706.07                Uncentered R2 =   0.8711
Residual SS             =  1755850.432                Root MSE      =     93.7

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1155622   .0057918    19.95   0.000     .1042105    .1269138
      kstock |   .2306785    .025284     9.12   0.000     .1811227    .2802342
       _cons |  -42.71437   9.440069    -4.52   0.000    -61.21656   -24.21217
------------------------------------------------------------------------------
Included instruments: mvalue kstock
------------------------------------------------------------------------------

. chatest, cluster(year)
Test of:
heteroskedasticity (nR2, homokurt assumed)=        91.790  Chi-sq( 5) p=0.0000
heteroskedasticity (no homokurt assumed)=          33.377  Chi-sq( 5) p=0.0000
heteroskedasticity & clustering=                   11.075  Chi-sq( 5) p=0.0499
heteroskedasticity & clustering (no centering)=    19.674  Chi-sq( 5) p=0.0014
het+clustering vs het-only (centering)=           211.206  Chi-sq( 5) p=0.0000
het+clustering vs het-only (no centering)=         12.588  Chi-sq( 5) p=0.0276
clustering vs homosked (no centering)=           2673.899  Chi-sq( 5) p=0.0000
het+clustering vs clust-only=                      19.294  Chi-sq( 5) p=0.0017
het+clustering vs clust-only (no centering)=       14.639  Chi-sq( 5) p=0.0120
```

```
. xtivreg2 invest mvalue kstock, fe
FIXED EFFECTS ESTIMATION
------------------------

Number of groups =        10                  Obs per group: min =        20
                                                             avg =      20.0
                                                             max =        20

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                Number of obs =       200
                                                F(  2,   188) =    309.01
                                                Prob > F      =    0.0000
Total (centered) SS     =  2244352.228          Centered R2   =    0.7668
Total (uncentered) SS   =  2244352.228          Uncentered R2 =    0.7668
Residual SS             =  523478.1139          Root MSE      =     52.49

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1101238   .0117941     9.34   0.000     .0870077    .1332399
      kstock |   .3100653   .0172629    17.96   0.000     .2762306       .3439
------------------------------------------------------------------------------
Included instruments: mvalue kstock
------------------------------------------------------------------------------

. xtchatest, cluster(company)
Test of:
heteroskedasticity (nR2, homokurt assumed)=        96.573  Chi-sq( 3) p=0.0000
heteroskedasticity (no homokurt assumed)=           4.306  Chi-sq( 3) p=0.2303
heteroskedasticity & clustering=                    8.303  Chi-sq( 3) p=0.0402
heteroskedasticity & clustering (no centering)=     3.672  Chi-sq( 3) p=0.2991
het+clustering vs het-only (centering)=             4.495  Chi-sq( 3) p=0.2128
het+clustering vs het-only (no centering)=          3.305  Chi-sq( 3) p=0.3470
clustering vs homosked (no centering)=              3.371  Chi-sq( 3) p=0.3379
het+clustering vs clust-only=                       3.315  Chi-sq( 3) p=0.3456
het+clustering vs clust-only (no centering)=        1.932  Chi-sq( 3) p=0.5867
```

Some empirical examples

Compare VCE estimates from a panel dataset computed under assumptions:

- *i.i.d.*
- HAC with 4 lags
- two-way cluster-robust HAC, 4 lags (correlated common shocks)

Table: Investment equation using grunfeld

	(1) iid	(2) hac4	(3) hac4_2way
mvalue	0.110***	0.110***	0.110***
	(0.0119)	(0.0238)	(0.00794)
kstock	0.310***	0.310***	0.310***
	(0.0174)	(0.0517)	(0.0344)
N	200	200	200

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table: Investment equation using grunfeld

	(1) iid	(2) hac4	(3) hac4_2way
mvalue	0.110***	0.110***	0.110***
	(0.0119)	(0.0238)	(0.00794)
kstock	0.310***	0.310***	0.310***
	(0.0174)	(0.0517)	(0.0344)
N	200	200	200

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Work in progress

We are currently working on the `chatest` and `xtchatest` routines in order to provide White-style tests for clustering vs. *i.i.d.*, and extending Kézdi's logic to two-way clustering.

The Stock–Watson heteroskedastic-robust VCE for the fixed effects estimator is already implemented in `xtivreg2` and will become a documented option.

We are also considering whether tests of this nature (which include White's (*Econometrica*, 1980) general test) may be adapted to consider only specific coefficients of interest. That is, are particular coefficients' standard errors and confidence intervals seriously affected by the assumed form of their VCE?