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 and CEPR

16th UK Stata Users Group Meeting, September 2010

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, one might naturally consider HAC standard errors: those robust to both heteroskedasticity and autocorrelation, familiar to economists as 'Newey–West' standard errors.

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, one might naturally consider HAC standard errors: those robust to both heteroskedasticity and autocorrelation, familiar to economists as 'Newey–West' standard errors.

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, 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.

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.

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.

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 ($g = 1, \ldots, G$).

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

Given the standard zero conditional mean assumption $E[u_{ig}|\mathbf{x}_{ig}] = 0$, 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?

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 ($g = 1, \ldots, G$).

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

Given the standard zero conditional mean assumption $E[u_{ig}|\mathbf{x}_{ig}] = 0$, 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 within-cluster correlation of errors can arise if the errors are not *i.i.d.*, but rather contain a common 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 random effects model of panel data, but may just as well arise in a cross-sectional context.

As in random effects, $Var[u_{ig}] = \sigma^2_\nu + \sigma^2_\zeta$ and $Cov[u_{ig}, u_{jg}] = \sigma^2_\nu, \forall i \neq j$.

Common shocks

The within-cluster correlation of errors can arise if the errors are not *i.i.d.*, but rather contain a common 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 random effects model of panel data, but may just as well arise in a cross-sectional context.

As in random effects, $Var[u_{ig}] = \sigma^2_\nu + \sigma^2_\zeta$ and $Cov[u_{ig}, u_{jg}] = \sigma^2_\nu, \ \forall i \neq j$.

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.

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.

In fact, if we had a dataset containing a number of equal-sized clusters, and regressors taking on constant values within those clusters, 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.

This model also has a parallel to panel data: it is the *between estimator* (`xtreg, be`) applied in the special case where $\mathbf{x}$ values do not differ within-panel.

In fact, if we had a dataset containing a number of equal-sized clusters, and regressors taking on constant values within those clusters, 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.

This model also has a parallel to panel data: it is the *between estimator* (`xtreg, be`) applied in the special case where **x** values do not differ within-panel.

If in contrast 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.

Moulton (*REStat*, 1990) 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.

If in contrast 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.

Moulton (*REStat*, 1990) 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

In the context of panel or longitudinal data, correlated errors naturally arise within each panel unit, so that each panel unit or individual can be considered as a cluster. With a time dimension, 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.

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.

Panel/longitudinal data

In the context of panel or longitudinal data, correlated errors naturally arise within each panel unit, so that each panel unit or individual can be considered as a cluster. With a time dimension, 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.

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.

In the context of a fixed-T, large N panel in which the common fixed-effects estimator (`xtreg, fe`) is applied, Stock and Watson (*Econometrica*, 2008) showed that the conventional 'robust' or sandwich VCE estimator is inconsistent if $T > 2$. This result applies even in the presence of serially uncorrelated errors. They present a bias-adjusted estimator that circumvents this problem, and illustrate how the asymptotically equivalent one-way cluster-robust VCE estimator we discuss next will provide consistent (although not fully efficient) estimates.

Given Stock and Watson's critique, Stata's `xtreg, fe` command was modified in version 10.0 (25feb2008) to automatically apply `vce(cluster id)`, where `id` is the panel unit identifier variable, if the `robust` option is specified.

In the context of a fixed-T, large N panel in which the common fixed-effects estimator (`xtreg, fe`) is applied, Stock and Watson (*Econometrica*, 2008) showed that the conventional 'robust' or sandwich VCE estimator is inconsistent if $T > 2$. This result applies even in the presence of serially uncorrelated errors. They present a bias-adjusted estimator that circumvents this problem, and illustrate how the asymptotically equivalent one-way cluster-robust VCE estimator we discuss next will provide consistent (although not fully efficient) estimates.

Given Stock and Watson's critique, Stata's `xtreg, fe` command was modified in version 10.0 (25feb2008) to automatically apply `vce(cluster id)`, where `id` is the panel unit identifier variable, if the `robust` option is specified.

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 White (*Asymptotic Theory for Econometricians*, 1984). The cluster-robust estimate takes the sandwich form

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

where

$$\hat{\Omega} = \sum_{g=1}^{G} \mathrm{X}_{\mathrm{g}}' \hat{u}_g \hat{u}_g' \mathrm{X}_{\mathrm{g}}$$

with $\hat{u}_g = \mathrm{y}_{\mathrm{g}} - \mathrm{X}_{\mathrm{g}}\hat{\beta}$ and g indicating membership in the g^{th} cluster.

This formula is derived from a more general specification where we consider the population moments, $E(x_i u_i)$, where u_i are the error terms. The corresponding sample moments are

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

where $\hat{u}_i$ are the residuals computed from point estimates $\hat{\beta}$.

The VCE of $\hat{\beta}$ is then

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

where $\hat{\Omega}$ is the estimated VCE of $\bar{g}(\hat{\beta})$, with its form based upon our assumptions about the properties of the error process u.
$\hat{Q}_{xx}^{-1}$ is merely $(\mathrm{X}'\mathrm{X})^{-1}$.

This formula is derived from a more general specification where we consider the population moments, $E(x_i u_i)$, where u_i are the error terms. The corresponding sample moments are

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

where $\hat{u}_i$ are the residuals computed from point estimates $\hat{\beta}$.

The VCE of $\hat{\beta}$ is then

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

where $\hat{\Omega}$ is the estimated VCE of $\bar{g}(\hat{\beta})$, with its form based upon our assumptions about the properties of the error process u.
$\hat{Q}_{xx}^{-1}$ is merely $(\mathrm{X}'\mathrm{X})^{-1}$.

Under the classical assumptions of *i.i.d.* errors: independence and conditional homoskedasticity, $\Omega = \sigma_u^2 \mathrm{I}$, and the estimate of V is merely $s^2(\mathrm{X}'\mathrm{X})^{-1}$, where s^2 is a consistent estimate of σ_u^2.

If we relax the assumption of conditional homoskedasticity,

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

with the VCE estimator as

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

the Huber–sandwich–White 'robust' estimator of the VCE, as invoked by the `robust` option in Stata. The expression for $\hat{\Omega}$ is a single sum over observations as we are maintaining the assumption of independence of each pair of errors.

Under the classical assumptions of *i.i.d.* errors: independence and conditional homoskedasticity, $\Omega = \sigma_u^2 \mathrm{I}$, and the estimate of V is merely $s^2(\mathrm{X}'\mathrm{X})^{-1}$, where s^2 is a consistent estimate of σ_u^2.

If we relax the assumption of conditional homoskedasticity,

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

with the VCE estimator as

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

the Huber–sandwich–White 'robust' estimator of the VCE, as invoked by the `robust` option in Stata. The expression for $\hat{\Omega}$ is a single sum over observations as we are maintaining the assumption of independence of each pair of errors.

In the cluster-robust case, where we allow for dependence between errors belonging to the same cluster, the VCE of $\bar{g}(\hat{\beta})$ becomes

$$\hat{\Omega} = \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i \sum_{i=1}^{N} \hat{g}_i'$$

where the double sum will include within-cluster cross products and between-cluster cross products of $(\mathbf{x}_i \hat{u}_i)$. By the assumption of independence across clusters, all terms involving errors in different clusters will be dropped, as they have zero expectation. The remaining within-cluster terms involve only the sums $(\mathbf{x}_i \hat{u}_i)$ for observations in each cluster, giving rise to the formula for $\hat{\Omega}$ which we presented earlier.

In the special case where errors are heteroskedastic but still independently distributed, the number of clusters G is equal to the number of observations, N, and each cluster is of size one. In this case the cluster-robust formula becomes the standard heteroskedasticity-robust 'White' formula implemented by Stata's `robust` option, as presented above.

When clusters with $N_g > 1$ are considered, the number of clusters G should be compared to the number of parameters to be estimated, k. The rank of $VCE(\hat{\beta})$ is at most G, as the 'meat' in the sandwich contains only G 'super-observations'. This implies that we cannot test more than G restrictions on the parameter vector, possibly invalidating a test for overall significance of the model, while tests on smaller subsets of the parameters are possible.

In the special case where errors are heteroskedastic but still independently distributed, the number of clusters G is equal to the number of observations, N, and each cluster is of size one. In this case the cluster-robust formula becomes the standard heteroskedasticity-robust ‘White’ formula implemented by Stata’s `robust` option, as presented above.

When clusters with $N_g > 1$ are considered, the number of clusters G should be compared to the number of parameters to be estimated, k. The rank of $VCE(\hat{\beta})$ is at most G, as the ‘meat’ in the sandwich contains only G ‘super-observations’. This implies that we cannot test more than G restrictions on the parameter vector, possibly invalidating a test for overall significance of the model, while tests on smaller subsets of the parameters are possible.

Bias in the cluster-robust estimator

While the formula for $\hat{\Omega}$ is appropriate as the number of clusters G goes to infinity, finite-sample corrections are usually applied to deal with downward bias in the cluster-robust standard errors. Stata uses $\sqrt{c}\hat{u}_g$ in computing $\hat{\Omega}$, with $c \simeq \frac{G}{G-1}$. 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 $\frac{N-1}{N-K}$ 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.

Bias in the cluster-robust estimator

While the formula for $\hat{\Omega}$ is appropriate as the number of clusters G goes to infinity, finite-sample corrections are usually applied to deal with downward bias in the cluster-robust standard errors. Stata uses $\sqrt{c}\hat{u}_g$ in computing $\hat{\Omega}$, with $c \simeq \frac{G}{G-1}$. 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 $\frac{N-1}{N-K}$ 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 $N \to \infty$. The approach that Stata follows involves using the t distribution with $G - 1$ degrees of freedom rather than $N - k$ 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_{G-L}, where L is the number of regressors constant within cluster, as an even more conservative approach.

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 $N \rightarrow \infty$. The approach that Stata follows involves using the t distribution with $G - 1$ degrees of freedom rather than $N - k$ 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_{G-L}, where L is the number of regressors constant within cluster, as an even more conservative approach.

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

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

As is well known from analysis of this model in the special case of longitudinal or panel data, the inclusion of the α_g parameters centers each cluster's residuals around zero. However, as the inclusion of these fixed-effect parameters does *nothing* to deal with potential intra-cluster correlation of errors, it is always advisable to question the *i.i.d.* error assumption and produce cluster-robust estimates of the VCE.

By analogy to the panel-data fixed effects model, we may note:

- The β parameters may be consistently estimated, but the coefficients of cluster-invariant regressors are not identified. For instance, if household data are clustered by state, state-level variables cannot be included.
- For $G \to \infty$, the α_g parameters cannot be consistently estimated due to the incidental parameter problem.
- The cluster-specific fixed effects, α_g, may be correlated with elements of $\mathbf{x}$.

By analogy to the panel-data fixed effects model, we may note:

- The β parameters may be consistently estimated, but the coefficients of cluster-invariant regressors are not identified. For instance, if household data are clustered by state, state-level variables cannot be included.
- For $G \rightarrow \infty$, the α_g parameters cannot be consistently estimated due to the incidental parameter problem.
- The cluster-specific fixed effects, α_g, may be correlated with elements of $\mathbf{x}$.

By analogy to the panel-data fixed effects model, we may note:

- The β parameters may be consistently estimated, but the coefficients of cluster-invariant regressors are not identified. For instance, if household data are clustered by state, state-level variables cannot be included.
- For $G \rightarrow \infty$, the α_g parameters cannot be consistently estimated due to the incidental parameter problem.
- The cluster-specific fixed effects, α_g, may be correlated with elements of $\mathbf{x}$.

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, as inference from such a model will be limited if $G < k$.

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.

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.

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.

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

$$\hat{\Omega} = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{I}(i,j) \left[\mathbf{x}_i \mathbf{x}_j' \, \hat{u}_i \hat{u}_j \right]$$

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

Computation of the two-way cluster-robust VCE is straightforward, as Thompson (SSRN WP, 2006) 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.

Computation of the two-way cluster-robust VCE is straightforward, as Thompson (SSRN WP, 2006) 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 G is too small to produce unbiased estimates. The theory underlying two-way clustering relies on asymptotics in the smaller number of clusters: that is, the dimension containing fewer clusters. The two-way clustering approach is thus most sensible if there are 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 $\frac{N-1}{N-K}$. In `ivreg2`, both adjustment factors are invoked with the `small` option.

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 $\frac{N-1}{N-K}$. 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.

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.

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.

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*, forthcoming; 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.

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*, forthcoming; 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.

Combining HAC and cluster-robust methods

In the context of panel data, the HAC (or 'Newey–West') estimator of the VCE allows for arbitrary serial correlation within each panel's errors, but the assumption of independence across units' errors is preserved. If we employ clustering by time period, we allow for common shocks across panel units' errors. When combined with the HAC estimator, this models common correlated shocks which will damp over time.

As the application of the HAC estimator requires a sufficient number of time periods per panel unit for consistency, the combination of HAC with clustering in the time dimension will be similarly demanding, and should not be employed in short panels.

Combining HAC and cluster-robust methods

In the context of panel data, the HAC (or ‘Newey–West’) estimator of the VCE allows for arbitrary serial correlation within each panel’s errors, but the assumption of independence across units’ errors is preserved. If we employ clustering by time period, we allow for common shocks across panel units’ errors. When combined with the HAC estimator, this models common correlated shocks which will damp over time.

As the application of the HAC estimator requires a sufficient number of time periods per panel unit for consistency, the combination of HAC with clustering in the time dimension will be similarly demanding, and should not be employed in short panels.

Combining GLS and cluster-robust methods

Cluster-robust techniques, as generalizations of the heteroskedasticity-robust 'White' VCE estimator, do not assume any particular form for the errors beyond the assumption of potential correlation within clusters. In some applications, we make use of generalized least squares (GLS) techniques to explicitly model departures from *i.i.d.* errors. For instance, we may explicitly deal with groupwise heteroskedasticity by estimating s^2 values from each group of errors, and use weights in `regress` to produce GLS estimates.

This specification still presumes independence among the errors. That assumption may be relaxed by considering the groups as clusters, and computing the weighted least squares regression with a cluster-robust VCE: a technique that could be extended to two-way or multi-way clustering.

Combining GLS and cluster-robust methods

Cluster-robust techniques, as generalizations of the heteroskedasticity-robust ‘White’ VCE estimator, do not assume any particular form for the errors beyond the assumption of potential correlation within clusters. In some applications, we make use of generalized least squares (GLS) techniques to explicitly model departures from *i.i.d.* errors. For instance, we may explicitly deal with groupwise heteroskedasticity by estimating s^2 values from each group of errors, and use weights in `regress` to produce GLS estimates.

This specification still presumes independence among the errors. That assumption may be relaxed by considering the groups as clusters, and computing the weighted least squares regression with a cluster-robust VCE: a technique that could be extended to two-way or multi-way clustering.

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 which considers the contrasts between an estimator of the VCE that is always consistent and one imposing more restrictive assumptions on the error process. A quadratic form in the vector of contrasts, in a framework similar to a Hausman test, yields a test statistic distributed χ^2 under the null hypothesis that the more restrictive assumptions (e.g., independence of errors, or *i.i.d.* errors) are supported.

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 which considers the contrasts between an estimator of the VCE that is always consistent and one imposing more restrictive assumptions on the error process. A quadratic form in the vector of contrasts, in a framework similar to a Hausman test, yields a test statistic distributed χ^2 under the null hypothesis that the more restrictive assumptions (e.g., independence of errors, or *i.i.d.* errors) are supported.

Kézdi's 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.

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 is implemented as Stata command `chatest`, with the panel counterpart `xtchatest`. These routines have not been released, as they are still under development. We hope to extend them to tests of two-way (or multi-way) clustering.

Kézdi's 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.

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 is implemented as Stata command `chatest`, with the panel counterpart `xtchatest`. These routines have not been released, as they are still under development. We hope to extend them to tests of two-way (or multi-way) clustering.

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

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

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

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

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
$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

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 (company and time)

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 (company and time)

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 (company and time)

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
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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)

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)

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)	(2)	(3)
	iid	hac4	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.

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?

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.

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?