Cluster-Robust Inference

Consider the linear regression model

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{1}$$

where the data have been divided into G disjoint clusters. This model can also be written as

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{u}_g, \quad g = 1, \ldots, G, \tag{2}$$

Cluster g has N_g observations, so that $N = \sum_{g=1}^{G} N_g$, and the vectors and matrices in (2) have N_g rows.

$\boldsymbol{u}_g$ and $\boldsymbol{u}_h$ are assumed independent for $g \neq h$, but with arbitrary patterns of heteroskedasticity and dependence within each cluster.

Clusters might correspond to classrooms, schools, families, villages, hospitals, firms, industries, years, cities, counties, states, or countries.

The OLS estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{y}_g, \tag{3}$$

where $\boldsymbol{y}$ stacks the $\boldsymbol{y}_g$, and $\boldsymbol{X}$ stacks the $\boldsymbol{X}_g$.

The random part of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{u}_g = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g, \quad \boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g. \tag{4}$$

Statistical properties of $\hat{\boldsymbol{\beta}}$ depend on those of the **score vectors** $\boldsymbol{s}_g$.

To obtain consistency and asymptotic normality, Djogbenou, MacKinnon, and Nielsen (JoE, 2019) requires $G \to \infty$ and imposes assumptions on the $\boldsymbol{s}_g$.

Ideally, G would be large, and the distributions of the $\boldsymbol{s}_g$ would be the same for all clusters.

The true variance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \boldsymbol{\Sigma}_g \Big) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \quad \boldsymbol{\Sigma}_g = \text{E}(\boldsymbol{s}_g \boldsymbol{s}_g^\top). \tag{5}$$

We need to estimate the $\boldsymbol{\Sigma}_g$, and this can be done in several ways.

The most popular **cluster-robust variance estimator**, or **CRVE**, is

$$\text{CV}_1\text{:} \qquad \frac{G(N-1)}{(G-1)(N-k)} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top \Big) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{6}$$

where $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g$ is the **empirical score vector** for cluster g.

Inference is usually based on the $t(G-1)$ distribution, although it is strictly valid only under extremely unrealistic assumptions. See Bester, Conley, and Hansen (JoE, 2011).

There are two well-known alternatives to CV_1:

$$\text{CV}_2: \qquad (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \grave{\boldsymbol{s}}_g \grave{\boldsymbol{s}}_g^\top \Big) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{7}$$

where $\grave{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{M}_{gg}^{-1/2} \hat{\boldsymbol{u}}_g$, and $\boldsymbol{M}_{gg} = \mathbf{I}_{N_g} - \boldsymbol{X}_g (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_g^\top$ is the g^{th} diagonal block of the projection matrix $\boldsymbol{M}_{\boldsymbol{X}}$, which satisfies $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$. $\boldsymbol{M}_{gg}^{-1/2}$ is its inverse symmetric square root.

$$\text{CV}_3: \qquad \frac{G-1}{G} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \acute{\boldsymbol{s}}_g \acute{\boldsymbol{s}}_g^\top \Big) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{8}$$

where $\acute{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{M}_{gg}^{-1} \hat{\boldsymbol{u}}_g$. CV_3 is actually a cluster jackknife estimator.

CV_2 and CV_3 were proposed in Bell and McCaffrey (SM, 2002). When $G = N$, they reduce to the familiar HC_2 and HC_3 estimators of MacKinnon and White (JoE, 1985).

Modeling Intra-Cluster Dependence

Intra-cluster correlations of the disturbances and regressors, and hence of the scores, can arise for many reasons.

The simplest and most popular model is the random-effects, or error-components, model

$$u_{gi} = \lambda \varepsilon_g + \varepsilon_{gi}, \tag{9}$$

where the idiosyncratic shock $\varepsilon_{gi} \sim \text{iid}(0, \omega^2)$ is independent of the cluster-wide shock $\varepsilon_g \sim \text{iid}(0, 1)$.

- For (9) the variance matrix $\mathbf{\Omega}_g$ has diagonal elements $\lambda^2 + \omega^2$ and off-diagonal elements λ^2.
- Within every cluster, the disturbances are equi-correlated, with correlation coefficient $\lambda^2 / (\lambda^2 + \omega^2)$.
- If we include cluster fixed effects, as is very commonly done, they absorb the ε_g and remove all intra-cluster correlation.

A slightly more complicated model is the **factor model**

$$u_{gi} = \lambda_{gi}\varepsilon_g + \varepsilon_{gi}. \tag{10}$$

The effect of ε_g on u_{gi} is given by a weight, or factor loading, λ_{gi}. The λ_{gi} could be either fixed parameters or random variables.

- For the factor model (10), $\mathrm{E}(u_{gi}) = 0$, and $\mathrm{Var}(u_{gi}) = \lambda_{gi}^2 + \omega^2$.
- Cluster dependence is characterized by $\mathrm{Cov}(u_{gi}, u_{gj}) = \lambda_{gi}\lambda_{gj}$, which differs across (i, j) pairs.
- $\mathrm{Cov}(u_{gi}, u_{gj})$ is zero only when the factor loadings are all zero and constant when they are all constant.

Consider a model for student achievement, where observations are for students, and clusters denote classrooms.

Then ε_{gi} measures unobserved student-specific characteristics, ε_g measures unobserved teacher quality, and λ_{gi} measures the extent to which student i is affected by teacher quality.

Including cluster fixed effects transforms the factor model (10) into

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g)\varepsilon_g + (\varepsilon_{gi} - \bar{\varepsilon}_g). \tag{11}$$

The averages are taken across observations within each cluster, so that $\bar{u}_g = N_g^{-1} \sum_{i=1}^{N_g} u_{gi}$, and likewise for the $\bar{\lambda}_g$ and the $\bar{\varepsilon}_g$.

The intra-cluster covariance for (11) is

$$\text{Cov}(u_{gi} - \bar{u}_g, u_{gj} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gj} - \bar{\lambda}_g), \tag{12}$$

which is zero if and only if, for each g, the λ_{gi} are the same for all i.

Including fixed effects almost always reduces intra-cluster correlations, but rarely will it eliminate them.

Only for (9), and for more general models where the λ_g and/or the ω_g vary across clusters, will it remove all intra-cluster dependence.

Especially when clusters are large, we generally need both cluster fixed effects and a CRVE.

At What Level Should We Cluster?

In many cases, there is more than one level at which we could cluster.

- With data on educational outcomes, we may be able to cluster by classroom, by school, or perhaps by school district.
- With data that are coded geographically, we may be able to cluster by county, by state, or even by region.

Suppose there are two possible levels of clustering, coarse and fine, with one or more fine clusters nested within each of the coarse clusters.

With G coarse clusters, the middle matrix in (5) is $\sum_{g=1}^{G} \mathbf{\Sigma}_g$. If each coarse cluster contains M_g fine clusters indexed by h, then

$$\mathbf{\Sigma}_g = \sum_{h_1=1}^{M_g} \sum_{h_2=1}^{M_g} \mathbf{\Sigma}_{g,h_1h_2}, \tag{13}$$

where $\mathbf{\Sigma}_{g,h_1h_2}$ denotes the covariance matrix of the scores for fine clusters h_1 and h_2 within coarse cluster g.

The difference between the middle matrices for coarse and fine clustering is

$$\sum_{g=1}^{G} \boldsymbol{\Sigma}_g - \sum_{g=1}^{G} \sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh} = 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2=h_1+1}^{M_g} \boldsymbol{\Sigma}_{g,h_1h_2}. \tag{14}$$

Under the assumption of fine clustering, the terms on the RHS of (14) are all zero. Under the assumption of coarse clustering, at least some of them are non-zero, and (14) must therefore be estimated.

- If we cluster at the fine level when coarse clustering is appropriate, the CRVE is inconsistent.
- If we cluster at the coarse level when fine clustering is appropriate, the CRVE has to estimate (14) even though it is actually zero.
- This causes loss of power and perhaps poor finite-sample inference, especially when G is small.

Rules of thumb for choosing the level at which to cluster:

- Cluster at the coarsest feasible level.
- Cluster at whichever level yields the largest standard error(s) for the coefficient(s) of interest.

When G is small, cluster-robust standard errors tend to be too small, perhaps much too small. Thus the first rule of thumb is dangerous.

However, the second rule of thumb may be too conservative.

- For treatment regressions, the level at which to cluster depends on the level at which treatment is applied. See Bertrand, Duflo, and Mullainathan (QJE, 2004) and the discussion on the next page.

It is possible to test the level of clustering, but pre-test issues arise.

- Ibragimov and Müller (REStat, 2016) proposes a test based on estimating the model separately for each cluster.
- MacKinnon, Nielsen, and Webb (JoE, 2023) proposes **score-variance tests** based on the empirical analogs of (14).

It treatment is applied *randomly* at some level, then we do not need to cluster at a higher level.

With random treatment at the individual level, we can use an HCCME.

With random treatment at, say, the firm level, then we must cluster at the firm level. But there should be no need to cluster at a higher level.

These results follow from the fact that the middle matrix in a CRVE is an estimate of the variance of the score vectors.

Each element of the score vector for the treatment dummy is a disturbance, say u_{gi}, times the treatment dummy projected off all the other regressors, say $v_{gi} = (\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{d})_{gi}$.

The covariance of $u_{gi}v_{gi}$ and $u_{hj}v_{hj}$ is

$$\mathrm{E}(u_{gi}u_{hj}v_{gi}v_{hj}) = \mathrm{E}(u_{gi}u_{hj})\mathrm{E}(v_{gi}v_{hj}). \tag{15}$$

So whenever $\mathrm{E}(v_{gi}v_{hj}) = 0$, the observation pair indexed by gi and hj does not contribute to $\boldsymbol{\Sigma}$.

Leverage and Influence

Heterogeneity across clusters makes asymptotic inference less reliable.

Classic measures of heterogeneity are **leverage** and **influence**.

These are generalized to cluster-level measures in MacKinnon, Nielsen, and Webb (SJ, 2023).

If estimates change a lot when a cluster is deleted, it is said to be **influential**. We should be wary of highly influential clusters.

To identify influential clusters, construct $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and $\boldsymbol{X}_g^\top \boldsymbol{y}_g$ for $g = 1, \ldots, G$. Then the estimates omitting cluster g are

$$\hat{\boldsymbol{\beta}}^{(g)} = (\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}_g^\top \boldsymbol{X}_g)^{-1} (\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}_g^\top \boldsymbol{y}_g). \tag{16}$$

We cannot partial out regressors other than cluster fixed effects prior to computing the $\hat{\boldsymbol{\beta}}^{(g)}$, because the latter would then depend indirectly on the observations for the g^{th} cluster.

When β_j is a parameter of particular interest, it is good to report the $\hat{\beta}_j^{(g)}$ for $g = 1, \ldots, G$ in either a histogram or a table.

If $\hat{\beta}_j^{(h)}$ differs a lot from $\hat{\beta}_j$ for some h, then cluster h is influential.

In some extreme cases, it may be impossible to compute $\hat{\beta}_j^{(h)}$ for some h. If so, the original estimates should probably not be believed. This will happen, for example, when cluster h is the only treated one.

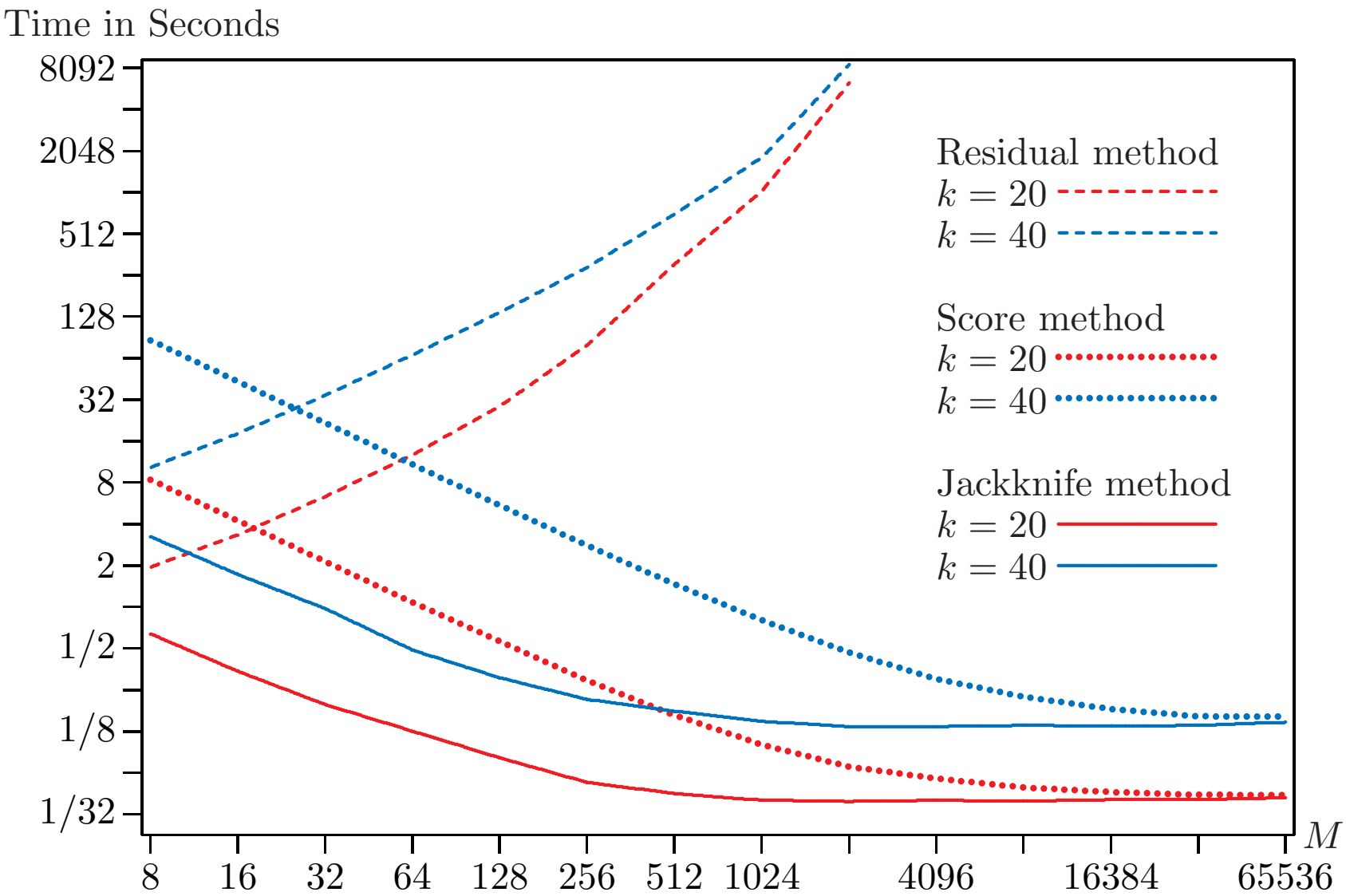
An alternative way to write CV_3 is

$$\mathrm{CV}_3\text{:} \qquad \frac{G-1}{G} \sum_{g=1}^{G} (\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^\top. \tag{17}$$

This is the matrix version of the classic jackknife variance estimator. Unless all clusters are tiny, (17) is much faster to compute than (8).

The `Stata` package `summclust` and the built-in option `vce(jackknife,mse)` calculate CV_3 standard errors based on (17). These can be much more reliable than CV_1 standard errors.

Figure 1: Timings for three ways to compute CV_3



A **high-leverage cluster** is one for which the regressors contain a lot of information about the fitted values.

High-leverage observations are associated with a high value of h_i, the i^{th} diagonal element of $\boldsymbol{H} = \boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

The analog of h_i in the cluster case is the $N_g \times N_g$ matrix $\boldsymbol{H}_g = \boldsymbol{X}_g(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}_g^\top$. Thus we can define

$$L_g = \text{Tr}(\boldsymbol{H}_g) = \text{Tr}\left(\boldsymbol{X}_g^\top \boldsymbol{X}_g(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right), \quad g = 1, \ldots, G. \tag{18}$$

The L_g are easy to compute because we have already calculated $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and the $\boldsymbol{X}_g^\top \boldsymbol{X}_g$.

For a cluster with one observation, L_g reduces to the usual measure of leverage at the observation level.

High-leverage clusters can be identified by comparing the L_g to their own average, which is k/G. If, for some h, L_h is substantially larger than k/G, then cluster h has high leverage.

A cluster can have high leverage either because N_h is much larger than G/N or because the matrix $\boldsymbol{X}_h$ is somehow extreme relative to the other $\boldsymbol{X}_g$ matrices, or both.

For example, L_h is likely to be much larger than k/G if cluster h is one of just a few treated clusters.

- When one of the regressors is a fixed-effect dummy for cluster g, the matrices $\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}_g^\top \boldsymbol{X}_g$ are singular.
- This problem solves itself if we partial out the fixed-effect dummies and replace $\boldsymbol{X}$ by $\tilde{\boldsymbol{X}}$ and $\boldsymbol{y}$ by $\tilde{\boldsymbol{y}}$, the matrix and vector of deviations from cluster means.
- The sum of the L_g is k. If we partial out l regressors, then the sum will instead be $k - l$.
- The gj^{th} element of $\tilde{\boldsymbol{y}}$ is $y_{gj} - N_g^{-1} \sum_{i=1}^{N_g} y_{gi}$, and similarly for the $\boldsymbol{X}_{gi}$. Since this depends only on observations for cluster g, the jackknife CV_3 estimator (17) remains valid.

Critical Values

It is common to use the $t(G-1)$ distribution for inference based on t statistics, and `Stata` does so by default.

Critical values can also be based on various approximations, which depend on $\boldsymbol{X}$ and an assumed form of $\boldsymbol{\Omega}$.

Bell and McCaffery (SM, 2002) suggests methods for CV_2 and CV_3 t-statistics based on the Student's t distribution with an estimated degrees-of-freedom (d-o-f) parameter.

- These employ a "Satterthwaite approximation" and calculate the d-o-f parameter under the assumption that $\text{Var}(\boldsymbol{u}) = \sigma^2\mathbf{I}$.
- The d-o-f parameter is different for every hypothesis to be tested, and it can be much less than $G-1$.

Imbens and Kolesár (REStat, 2016) proposes a similar procedure for t-tests based on CV_2 under the assumption that $\text{Var}(\boldsymbol{u})$ corresponds to a cluster random-effects model. But this makes no sense if there are cluster fixed effects.

Young (2016) proposes a related method that uses CV_1 instead of CV_2.

Pustejovsky and Tipton (JBES, 2018) generalizes the procedure of B&M (2002) to Wald tests based on CV_2.

- Simulations suggest that their Wald tests rarely over-reject but often under-reject, sometimes quite severely.

Very recently, Hansen (WP, 2024) proposes a method for inference based on CV_3 t statistics, again under the assumption that $\mathbf{\Omega} = \sigma^2\mathbf{I}$.

It involves estimating two parameters. One of them shrinks the CV_3 standard error, and the other is a d-o-f parameter, smaller than $G-1$ and sometimes very small.

The Hansen procedure seems to avoid the serious under-rejection that can occur if we combine CV_3 with the $t(G-1)$ distribution.

Hansen has a `Stata` package called `jregress`, which runs an OLS regression, computes his modified CV_3 standard errors, and uses them and the t distribution to compute P values and confidence intervals.

Two-Way Clustering

There can be clustering in two or more dimensions. For example,

- There may be clustering by jurisdiction and also by time period.
- In finance, there is often clustering by both firm and year.

Thus, instead of (2), we might have

$$\boldsymbol{y}_{gh} = \boldsymbol{X}_{gh}\boldsymbol{\beta} + \boldsymbol{u}_{gh}, \quad g = 1, \ldots, G, \quad h = 1, \ldots, H, \tag{19}$$

where $\boldsymbol{y}_{gh}$, $\boldsymbol{u}_{gh}$, and $\boldsymbol{X}_{gh}$ contain the rows of $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{X}$ for cluster g in the first clustering dimension and cluster h in the second.

The GH clusters into which the data are divided in (19) represent the intersection of the two clustering dimensions. Note that N_{gh} may be 0.

If there are N_g observations in cluster g, N_h observations in cluster h, and N_{gh} observations in cluster gh, then

$$N = \sum_{g=1}^{G} N_g = \sum_{h=1}^{H} N_h = \sum_{g=1}^{G} \sum_{h=1}^{H} N_{gh}. \tag{20}$$

The scores for the clusters in the first and second dimensions are $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g$ and $\boldsymbol{s}_h = \boldsymbol{X}_h^\top \boldsymbol{u}_h$, and for the intersections $\boldsymbol{s}_{gh} = \boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh}$.

If we assume that

$$\begin{aligned} \boldsymbol{\Sigma}_g &= \mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_g^\top),\ \boldsymbol{\Sigma}_h = \mathrm{E}(\boldsymbol{s}_h \boldsymbol{s}_h^\top), \\ \boldsymbol{\Sigma}_{gh} &= \mathrm{E}(\boldsymbol{s}_{gh} \boldsymbol{s}_{gh}^\top),\ \mathrm{E}(\boldsymbol{s}_{gh} \boldsymbol{s}_{g'h'}^\top) = \boldsymbol{0} \text{ for } g \neq g' \text{ and } h \neq h', \end{aligned} \tag{21}$$

then the variance matrix of the scores is seen to be

$$\boldsymbol{\Sigma} = \sum_{g=1}^{G} \boldsymbol{\Sigma}_g + \sum_{h=1}^{H} \boldsymbol{\Sigma}_h - \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}. \tag{22}$$

The scores are assumed to be independent whenever they do not share a cluster along either dimension.

It is important to distinguish between two-way clustering and clustering by the intersection of the two dimensions.

With clustering by intersection, all three terms on the r.h.s. of (22) would be equal, so that $\boldsymbol{\Sigma} = \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}$. **Extremely restrictive!**

An estimator of the variance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{\Sigma}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1},$$
$$\hat{\boldsymbol{\Sigma}} = \sum_{g=1}^{G} \hat{\mathbf{s}}_g \hat{\mathbf{s}}_g^\top + \sum_{h=1}^{H} \hat{\mathbf{s}}_h \hat{\mathbf{s}}_h^\top - \sum_{g=1}^{G} \sum_{h=1}^{H} \hat{\mathbf{s}}_{gh} \hat{\mathbf{s}}_{gh}^\top. \tag{23}$$

Each of the matrices on the r.h.s. of second equation in (23) is usually multiplied by a scalar factor to correct for degrees of freedom.

The third term in (22) is subtracted to avoid double counting. Thus $\hat{\boldsymbol{\Sigma}}$ may not always be positive definite. What should we do?

- Use eigenvalue decomposition to force matrix to be positive semidefinite; see Cameron, Gelbach, and Miller (JBES, 2011).
- Omit the third term. But this can lead to inconsistency.
- Use largest standard error (smallest test statistic) from $\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\Sigma}}_G, \hat{\boldsymbol{\Sigma}}_H$.

See MacKinnon, Nielsen, and Webb (JBES, 2021; WP 1516, 2024), Davezies, D'Haultfoeuille and Guyonvarch (AS, 2021).