Recent Developments in Cluster-Robust Inference

Colin Cameron
Univ. of California - Davis

Joint work with Douglas L. Miller, Cornell UIniversity.

.

Webinar to International Association for Applied Econometrics
Detailed references are at https://appliedeconometrics.org/iaae-webinars

December 15, 2021

1. Introduction

- Cluster error correlation
  - errors are correlated within cluster (or group)
  - and independent across clusters
    - in the simplest case of one-way clustering.
- Many applications in microeconometrics have cluster error correlation.
- Erroneously assuming error independence can lead to wildly under-estimated standard errors
  - e.g. one-third of correct standard error.
- The standard cluster-robust inference methods
  - are valid asymptotically
  - but in many, many applications the asymptotics have not kicked in
    - tests over-reject and confidence intervals undercover
    - called the "few clusters" problem but can occur with many clusters.

- Surveys are
  - A. Colin Cameron and Douglas L. Miller (2015), "A Practitioner's Guide to Robust Inference with Clustered Data," *Journal of Human Resources*, Spring 2015, Vol.50(2), pp.317-373.
  - James G. MacKinnon, Morten Ø. Nielsen, and Matthew D. Webb (2021), "Cluster-robust inference: A guide to empirical practice", QED Working Paper No. 1456.
- Recent texts place more emphasis on cluster-robust methods
  - Bruce E. Hansen (2022), *Econometrics*, Princeton University Press, forthcoming.
  - A. Colin Cameron and Pravin K. Trivedi (2022), *Microeconometrics using Stata*, Second edition, Stata Press, forthcoming.

Outline

2. Basics of Cluster-robust inference

- Linear model for G clusters with N_g individuals per cluster

$$\begin{aligned} y_{ig} &= \mathbf{x}'_{ig}\boldsymbol{\beta} + u_{ig}, i = 1, ..., N_g, \ g = 1, ..., G, \ N = \textstyle\sum_{g=1}^{G} N_g \\ \mathbf{y}_g &= \mathbf{X}'_g\boldsymbol{\beta} + \mathbf{u}_g, \quad g = 1, ..., G \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_g \end{aligned}$$

- Clustered errors: u_{ig} independent over g and correlated within g

$$\mathrm{E}[u_{ig}u_{jg'}|\mathbf{x}_{ig}, \mathbf{x}_{jg'}] = 0, \text{ unless } g = g'.$$

- Then OLS estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has

$$\begin{aligned} \mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathrm{E}[\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}|\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\textstyle\sum_{g=1}^{G} \mathrm{E}[\mathbf{X}'_g\mathbf{u}_g\mathbf{u}'_g\mathbf{X}_g|\mathbf{X}])(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

2.1 Cluster-robust variance matrix estimate

- For OLS with independent clustered errors

$$\mathrm{Var}[\widehat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}(\textstyle\sum_{g=1}^{G} \mathrm{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g|\mathbf{X}])(\mathbf{X}'\mathbf{X})^{-1}$$

- A (heteroskedastic- and) cluster-robust variance estimate (CRVE) is

$$\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}(\textstyle\sum_{g=1}^{G} \mathbf{X}_g'\widetilde{\mathbf{u}}_g\widetilde{\mathbf{u}}_g'\mathbf{X}_g)(\mathbf{X}'\mathbf{X})^{-1}.$$

- $\widetilde{\mathbf{u}}_g$ is a finite-sample correction to $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g'\widehat{\beta}$
  - Stata uses $\widetilde{\mathbf{u}}_g = \sqrt{c}\widehat{\mathbf{u}}_g$ where $c = \frac{G}{G-1} \times \frac{N-1}{N-K} \simeq \frac{G}{G-1}$.
- Stata: `vce(cluster)` option or `vce(robust)` option following `xtset`
- R: `sandwich` package CR1.

2.2 Two Different settings

- Setting 1: Individual in regions or schools or ... ("Moulton")
  - natural starting point is equicorrelated errors or exchangeable errors within cluster (e.g. random effects model $u_{ig} = \alpha_g + \varepsilon_{ig}$)
  - error correlation within cluster does not disappear with separation of observations
    - ★ marginal information contribution of an additional observation in a cluster can be very low.
- Setting 2: Panel data ("BDM")
  - now the individual unit is the cluster g (and i is time)
  - natural starting point is autocorrelated error within cluster
  - error correlation within cluster disappears with separation of observations.
- These different settings can lead to different asymptotic theory.

- The CR variance matrix estimate was proposed by
  - White (1984, book) for balanced case
  - Liang and Zeger (1986, *JASA*) for grouped data (biostatistics)
  - Arellano (1987, *JE*) for FE estimator for short panels.
- Asymptotic theory initially had fixed and constant N_g and $G \longrightarrow \infty$
- Subsequent theory allows various rates for N_g and G
  - Christian Hansen (2007, *JE*) for panel data also allows $T \longrightarrow \infty$
  - Carter, Schnepel and Steigerwald (2017, *REStat*) also allows $N_g \longrightarrow \infty$
  - Djogbenou, MacKinnon and Nielsen (2019, *JE*) and Bruce Hansen and Seojeong Lee (2019, *JE*)
    - ★ more general conditions with considerable cluster-size heterogeneity and normalization more complex than $\sqrt{G}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.
- Inclusion of fixed effects
  - in practice still leaves considerable within cluster correlation
  - can complicate proofs beyond one-way cluster for OLS.

2.3 Confidence Intervals and Hypothesis Tests

- For a single coefficient β, asymptotic theory gives

$$\frac{\widehat{\beta}-\beta_0}{\sqrt{\text{Var}[\widehat{\beta}]}} \sim N[0,1].$$

- In practice we need to replace $\text{Var}[\widehat{\beta}]$ with $\widehat{\text{V}}_{\text{CR}}[\widehat{\beta}]$.
- Standard ad hoc adjustment is to then use the $T(G-1)$ distribution

$$\frac{\widehat{\beta}-\beta_0}{\text{se}_{CR}[\widehat{\beta}]} \sim T(G-1).$$

- The $T(G-1)$ distribution has fatter tails and is better than $N[0,1]$
  - ad hoc though Bester, Conley and Hansen (2009, *JE*) derive for fixed-G asymptotics and dependent data with homogeneous $\mathbf{X}_g'\mathbf{X}_g$.
- But in practice with finite G, tests based $T(G-1)$ over-reject
  - and confidence intervals undercover.

2.4 Survey methods

- Complex survey data are clustered, stratified and weighted.
- The loss of efficiency due to clustering is called the design effect.
- Survey software controls for all three
  - e.g. Stata `svy` commands.
- Econometricians
  - 1. Get standard errors that cluster on PSU or higher
  - 2. Ignore stratification (with slight loss in efficiency)
  - 3. Sometimes weight and sometimes not.

3. Better One-way Cluster-Robust Inference

- Consider two-sided symmetric t-test

$$t = \frac{\widehat{\beta} - \beta_0}{\text{se}(\widehat{\beta})} \text{ has c.d.f } F(t)$$
$$p = 2 \times (1 - \widehat{F}^{-1}(|\widehat{t}|)$$

- Three primary challenges to obtaining correct inference
  - $\text{se}(\widehat{\beta})$ has large-cluster bias
  - $\text{se}(\widehat{\beta})$ has finite-cluster bias
  - $\text{se}(\widehat{\beta})$ is a noisy estimate of St.Dev.$[\widehat{\beta}]$
- Failure to adequately control for these challenges can make $\widehat{F}(t)$ a poor approximation for $F(t)$.
- Similar issues for confidence interval.

3.1 Challenge 1: Large-cluster bias in standard error

- First-order reason for clustering standard errors.
- Appropriate clustering gives valid inference for $G = \infty$.
- For one-way clustering the key is determining level to cluster at
  - e.g. with individual panel data: individual (?), household (?), state (?)
  - e.g. in early work many clustered on state-year pair rather than state.
- Trade-off: clustering at a broader level makes for noisier se$(\widehat{\beta})$ and is more likely to lead to "few" clusters.
- In some applications need more general clustering than one-way
  - Multi-way clustering
  - Dyadic clustering
  - Spatial correlation.

Design-based inference

- Standard inference uses randomness due to a model error u.
- Design-based inference for RCTs instead has randomness coming from treatment assignment.
- Abadie, Athey, Imbens, Wooldridge (2017, NBER WP) "When Should You Adjust Standard Errors for Clustering".
- Sampling where a subset of clusters are sampled randomly from a population of clusters.
- If all units in a given cluster get the same treatment
  - use cluster-robust standard errors.
- In other cases cluster-robust standard errors may be conservative.
- Su and Ding (2021, JRSSB) use designed-based inference approach for cluster-randomized experiments with regression ("model-assisted") based on cluster averages and on individual data.

3.2 Challenge 2: Small-cluster bias in standard error

- Parameter estimates $\widehat{\beta}$ overfit the data at hand.
- So residuals $\widehat{u}$ are always in some sense smaller on average than model errors u.
- Plugging $\widehat{u}$ into CRVE formula will produce $se(\widehat{\beta})$ that is too small
  - this problem goes away as $G \rightarrow \infty$.
- In heteroskedastic errors case this leads to HC2 and HC3 standard errors (MacKinnon and White (1985, *JE*)).
- Can generalize HC2 and HC3 to one-way cluster robust (Bell and McCaffrey 2002)
  - CR2 adjusts for leverage and CR3 is a jackknife.
  - most studies use CR1 (the Stata and R default).

Reasons for small-cluster bias in standard error

- Few clusters
  - G small
- When clusters are asymmetric
  - N_g varies across g
  - weights vary across g (if weighted LS)
  - design matrix $\mathbf{X}_g'\mathbf{X}_g$ varies across g
    - leading example is few treated clusters
  - $\Omega_g = E[\mathbf{u}_g'\mathbf{u}_g|\mathbf{X}_g]$ varies across g
  - interaction between Ω_g and $\mathbf{X}_g'\mathbf{X}_g$
- Typically: the larger and higher leverage clusters will be more over-fit.

Leverage and Influential Observations

- MacKinnon, Nielsen and Matthew D. Webb (2021, Sections 7 and 8) present and illustrate
  - cluster leverage measures based on $\mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$
  - cluster influence measures based on $\widehat{\beta}_{(g)}$ that omits cluster G
- MacKinnon, Nielsen and Matthew D. Webb (2022)
  - Stata `summclust` command for cluster leverage and influence.
- Young (2019, *QJE*) shows that leverage can lead to great over-rejection using the conventional CRVE.

3.3 Challenge 3: noise in standard error

- The noise in the standard error leads to distribution other than $N(0,1)$ with finite number of clusters.
- There are many suggested methods detailed below
  - use $T(G-1)$ as statistical packages do
  - use $t(G^*)$ where data-determined G^* is better than $G-1$
  - use a better distribution than $t(G^*)$
  - use a bootstrap with asymptotic refinement
  - use asymptotics with G fixed and $N_g \to \infty$
  - use randomization inference
  - use feasible GLS.

3.3.1 T with Different Degrees of freedom

- Imbens and Kolesar (2016, *REStat*).
    - Data-determined number of degrees of freedom for t and F tests
    - Builds on Satterthwaite (1946) and Bell and McCaffrey (2002).
    - Assumes normally distributed equicorrelated errors and uses CR2.
    - Match first two moments of test statistic with first two moments of χ^2.
    - $v^* = (\sum_{j=1}^{G} \lambda_j)^2 / (\sum_{j=1}^{G} \lambda_j^2)$ and λ_j are the eigenvalues of the $G \times G$ matrix $\mathbf{G}'\widehat{\Omega}\mathbf{G}$.
- Pustejovsky and Tipton (2017, *JBES*)
    - Extend Imbens and Kolesar to joint hypothesis tests.

T with Different Degrees of freedom (continued)

- Carter, Schnepel and Steigerwald (2017, *REStat*)
  - consider unbalanced clusters due to variation in N_g, variation in $\mathbf{X}_g$ and variation in Ω_g across clusters
  - provide asymptotic theory
  - propose a measure G^* of the effective number of clusters
  - that is data-determined aside from $\Omega_g = E[\mathbf{u}_g \mathbf{u}_g' | \mathbf{X}]$.
  - no proof that one should use $T(G^*)$ but it seems better than $T(G-1)$.
- Lee and Steigerwald (2018, *SJ*)
  - provide Stata add-on command `clusteff` that computes G^*
  - default is conservative as it assumes perfect within cluster correlation of errors
  - option `covariance()` allows specifying $\rho < 1$ with equicorrelated errors.

3.3.2 Exact Distribution

- Meiselman (2021, UT-Austin WP)
  - fixed effects model
  - assumes normally distributed equicorrelated errors
  - derives exact c.d.f. of t^2.

3.4 Cluster Bootstrap with Asymptotic Refinement

- There are several ways to bootstrap
  - different resampling methods
  - different ways to then use for inference
    - ★ in some cases can get an asymptotic refinement.
- A fairly general procedure to get an asymptotic refinement is
  - percentile-t bootstrap that bootstraps the t statistic
  - with cluster-pairs resampling that resamples with replacement $(\mathbf{y}_g, \mathbf{X}_g)$.
- Cameron, Gelbach and Miller (2008) in simulations find better performance with finite G if instead
  - resample residuals $\widehat{\mathbf{u}}_g$ holding $\mathbf{X}_g$ fixed ("wild" cluster bootstrap)
  - impose H_0 in getting the residuals.

Wild Restricted Cluster Bootstrap

1. Obtain the restricted LS estimator $\widehat{\boldsymbol{\beta}}$ that imposes H_0.
   Compute the residuals $\widehat{\mathbf{u}}_g$, $g = 1, ..., G$.
2. Do B iterations of this step. On the b^{th} iteration:
   1. For each cluster $g = 1, ..., G$:
      Form $\widehat{\mathbf{u}}_g^* = d_g \times \widehat{\mathbf{u}}_g$ where $d_g = -1$ or 1 each with probability 0.5
      Hence form $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g' \widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}_g^*$.
      This yields wild cluster bootstrap resample $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$.
   2. Calculate the OLS estimate $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$.
      Hence form the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \widehat{\beta}_1)/s_{\widehat{\beta}_{1,b}^*}$.
3. Reject H_0 at level α if and only if

$$w < w^*_{[\alpha/2]} \text{ or } w > w^*_{[1-\alpha/2]},$$

where $w^*_{[q]}$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

Wild Restricted Cluster Bootstrap (continued)

- Implementation is fast and easy for practitioners.
- Roodman, MacKinnon, Nielsen and Webb (2019, *SJ*)
  - `boottest` add-on command to Stata is very fast
  - implements wild and score bootstrap of Wald or score test for many estimators
  - provides confidence intervals by test inversion.
- MacKinnon (2022, E&S)
  - further computational savings using sums of products and cross-products of observations within each cluster.

Wild Restricted Cluster Bootstrap (continued)

- Webb (2014, QED WP 1315) proposed a 6-point distribution for d_g in $\widehat{\mathbf{u}}_g^* = d_g \widehat{\mathbf{u}}_g$
  - better when $G < 10$.
- MacKinnon and Webb (2017, *JAE*)
  - unbalanced cluster sizes worsens poor test size using $\mathrm{V}_{CR}[\widehat{\beta}]$.
  - wild cluster bootstrap does well.
- Djogbenou, MacKinnon, Nielsen (2019, *JE*)
  - prove that the Wild cluster bootstrap provides an asymptotic refinement (using Edgeworth expansions).
- Canay, Santos and Shaikh (2021, *REStat*)
  - provides randomization inference theory for the wild bootstrap when $N_g \to \infty$ and symmetry holds
  - considers both studentized and unstudentized test statistics.

3.5 Few treated clusters

- Few treated clusters
  - often arises especially in differences-in-differences settings
  - basic cluster-robust inference can work poorly.
- MacKinnon and Webb (2018, *PM*)
  - extreme problem if only one treated cluster as then the OLS residuals in that cluster sum to zero
  - this leads to too small a variance estimate.
- Solutions often require strong assumptions such as
  - exchangeability within cluster
  - homogeneity across cluster
  - symmetry
  - identification can be obtained using only within-cluster estimates.

Few treated clusters (continued)

- Wild cluster bootstrap with few (treated) clusters
  - MacKinnon and Webb (2018, *EJ*)
- T distribution for t statistics from cluster-level estimates
  - Ibragimov and Müller (2010, *JBES*)
    - ★ only within-group variation is relevant, separately estimate $\widehat{\boldsymbol{\beta}}_g s$ and average, G small and $N_g \rightarrow \infty$.
    - ★ rules out $y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + \mathbf{z}'_g\boldsymbol{\gamma} + u_{ig}$.
  - Ibragimov and Müller (2016, *REStat*)
    - ★ extend to allow treated and untreated groups.
- Difference in difference settings
  - Conley and Taber (2011) assume exchangeability and have fixed T, fixed treated clusters, number of control clusters $\rightarrow \infty$
  - Ferman and Pinto (2019) extend this to (known) heteroskedastic errors.

3.6 Randomization inference

- A permutation test (Fisher) provides a test of exact size.
- For settings where data are exchangeable under the null hypothesis
  - e.g. two-sample difference in means test with two samples from the same distribution
- The procedure:
  - 1. Compute the test statistic using the original sample.
  - 2. Recompute this test statistic for every permutation of the data.
  - 3. p-value $=$ fraction of times permuted test statistic $\geq$ original sample test statistic.

Randomization inference (continued)

- Extends to a regressor of interest is uncorrelated with other regressors
  - e.g. if the regressor is a randomly assigned treatment.
- Young (2019, *QJE*) does this and compares to conventional methods and bootstrap.
- MacKinnon and Webb (2020, *JE*) consider when treatment is not randomly assigned.
- MacKinnon and Webb (2019, book chapter) adjust when there are few possible randomizations.

Randomization inference (continued)

- Canay, Romano and Shaikh (2017, *Ecta*)
  - extend to symmetric limiting distribution of a function of the data under H_0
  - covers DinD with few clusters and many observations per cluster.
- Cai, Kim and Shaikh (2021)
  - Stata and R packages to implement in linear models with few clusters.
- Hagemann (2019, *JE*)
  - assigns placebo treatments to untreated clusters to get nearly exact sharp test of no effect of a binary treatment.
- Hagemann (2020)
  - a rearrangement test for a single treated cluster with a finite number of heterogeneous clusters.
- Hagemann (2021)
  - adjusts permutation inference to get non-sharp test on binary treatment with finitely many heterogeneous clusters.

4. Beyond One-way Clustering

- Richer forms of clustering than one-way
  - Multi-way clustering
  - Dyadic clustering
  - Spatial correlation.

4.1 Multi-way Clustering

- What if have two non-nested reasons for clustering
  - e.g. regress individual wages on job injury rate in industry and on job injury rate on occupation
  - e.g. matched employer - employee data.
- Obtain three different cluster-robust "variance" matrices by
  - cluster-robust in (1) first dimension, (2) second dimension, and (3) intersection of the first and second dimensions
  - add the first two variance matrices and, to account for double-counting, subtract the third.

$$\widehat{\mathrm{V}}_{\text{two-way}}[\widehat{\boldsymbol{\beta}}] = \widehat{\mathrm{V}}_{G}[\widehat{\boldsymbol{\beta}}] + \widehat{\mathrm{V}}_{H}[\widehat{\boldsymbol{\beta}}] - \widehat{\mathrm{V}}_{G\cap H}[\widehat{\boldsymbol{\beta}}]$$

- A simpler more conservative estimate drops the third term
  - this guarantees that $\widehat{\mathrm{V}}_{\text{two-way}}[\widehat{\boldsymbol{\beta}}]$ is positive definite.

Multi-way Clustering (continued)

- Independently proposed by
  - Cameron, Gelbach, and Miller (2006; 2011, *JBES*) in econometrics
  - Miglioretti and Heagerty (2006, *AJE*) in biostatistics
  - Thompson (2006; 2011, *JFE*) in finance
  - Extends to multi-way clustering.
- Davezies, D'Haultfoeuille and Guyonvarch (2021, *AS*)
  - provides empirical process theory that assumes exchangeability and propose a pigeonhole bootstrap.
- Menzel (2021, *Ecta*)
  - provides theory and proposes a bootstrap.
- MacKinnon, Nielsen and Matthew D. Webb (2021, JBES)
  - provide theory and propose various Wild bootstraps.
- Chiang, Kato and Sasaki (2021, *JASA*)
  - inference and bootstraps for exchangeable arrays.

- Villacorta (2017, WP)
  - proposes an improvement on 2-way cluster-robust for panel data when N and T are small
  - does FGLS using a spatial autoregressive model.
- Chiang, Hansen and Sasaki (2021, in preparation)
  - for panel data two-way controls for cluster dependence within i and within t
  - this paper adds two terms to control for serial dependence in common time effects.
- Powell (2020, WP) for panel data allows correlation across clusters.
- Chiang, Kato, Ma and Sasaki (2022, *JBES*)
  - multiway cluster-robust double/debiased machine learning.
- Verdier (2020, *REStat*)
  - linear model with two-way fixed effects and sparsely matched data.

4.2 Dyadic Clustering

- A dyad is a pair. An example is country pairs.
- The errors for two pairs are correlated with each other if they have one person in common.
  - Call the pairs (g, h) and (g', h')
  - Two-way picks up error correlation for cases with $g = g$ and $h = h'$
  - Dyadic-robust additionally picks up $g = h'$ and $h = g'$.
- Fafchamps and Gubert (2007, *JDE*)
  - provide variance matrix
  - apply to a sparse network where it makes little difference.
- Cameron and Miller (2014, WP)
  - apply to international trade data where the network is dense and find it makes a big difference.

Dyadic Clustering (continued)

- Aronow and Assenova (2015, Political Analysis)
  - prove variance estimate but not asymptotic normal distribution.
- Tabord-Meehan (2018, *JBES*)
  - use a central limit theorem for dependency graphs (S. Jannson (1988)).
- Graham, Niu and Powell (2019, WP)
  - consider kernel density estimation for undirected dyadic data
  - obtain variance estimator and asymptotic normal distribution.

4.3 Spatial Correlation

- Consider state-year panel data.
- Cluster assumes independence across states.
- Spatial correlation allows some dependence across states that decays with distance.
- Different asymptotics that uses mixing conditions.
- Driscoll and Kraay (1998, *REStat*) panel data when $T \longrightarrow \infty$
  - generalizes HAC to spatial correlation for panel data with $T \rightarrow \infty$.
- Cao, Christian Hansen, Kozbur and Villacorta (2021)
  - inference for dependent data with learned clusters.

5. Estimators other than OLS and FGLS

- The asymptotic cluster robust inference methods for OLS extend to other standard estimators
  - FGLS
  - linear IV
  - nonlinear m-estimator
  - GMM
  - quantile
- More challenging for these are
  - finite-cluster corrections
    - ★ e.g. Wild cluster bootstrap with refinement uses a residual
  - handling fixed effects.

5.1 Feasible GLS

- Potential efficiency gains for feasible GLS compared to OLS.
- For one-way clustering the feasible GLS estimator has

$$\widehat{V}_{\mathrm{CR}}[\widehat{\beta}_{\mathrm{FGLS}}] = \left(\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\Omega}_g^{-1}\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\widehat{\Omega}_g^{-1}\mathbf{X}_g\right)\left(\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X}\right)^{-}$$

- Stata offers many FGLS estimators with CR standard errors.
- Yet this is not done much in economics.
- Brewer and Crossley (2018, *JEM*)
  - panel data with fixed effects and AR(2) error and bias-adjust
  - find much better test size performance using BDM data.

5.2 Instrumental Variables

- Cluster-robust variance generalizes immediately.
  - main focus is on cluster-robust inference with weak instruments.
- Chernozhukov and Hansen (2008, *EL*)
  - Cluster-robust version of Anderson-Rubin test is immediate.
- Weak instruments diagnostics
  - First-stage F-statistic should be cluster-robust
- Olea and Pfleuger (2013, *JBES*)
  - a cluster-robust version of the Stock-Yogo relative asymptotic bias test.
- Magnusson (2010, *EJ*)
  - weak-instrument-robust tests and confidence intervals for IV estimation of linear, probit and tobit models
  - includes cluster-robust and two-way robust for not just AR.
- Finlay and Magnusson (2019, JAE)
  - residual and Wild cluster bootstraps for IV with weak instruments.
- Young (2021) considers leverage and clustering in applications.

5.3 Nonlinear m-estimators

- Cluster-robust methods extend to nonlinear estimators
  - e.g. logit and nonlinear GMM.
  - e.g. generalized estimating equations (Liang and Zeger 1986).
- Kline and Santos (2012, *EM*)
  - wild score bootstrap
  - rather than resample $\widehat{\mathbf{u}}_g$ resample the score $\mathbf{X}_g'\widehat{\mathbf{u}}_g$
  - this extends to nonlinear models such as logit and probit.

5.4 GMM

- Cluster-robust extends to GMM.
- Hansen and Lee (2019, *JE*)
  - provide very general asymptotic theory for clustered samples
- Hansen and Lee (2021, *Ecta*)
  - inference for Iterated GMM under misspecification
  - consider heteroskedastic errors (journal dropped clustering).
- Hansen and Lee (2020, WP)
  - also has clustered errors.
- Hwang (2019, *JE*)
  - two-step GMM fixed-G asymptotics with recentering of the CRVE used at the second step.

5.5 Quantile

- Parente and Silva (2016, *JEM*)
  - quantile regression with clustered data.
- Yoon and Galvao (2020, *QE*)
  - cluster-robust inference for panel quantile regression models with individual fixed effects and serial correlation.
- Hagemann (2017, *JASA*)
  - Cluster-robust bootstrap inference.

6. Conclusion

- Where clustering is present it is important to control for it.
- Most work is for OLS and one-way clustering.
- Even in this case it is not clearly established what is the best method when there are few clusters or clusters are very unbalanced / heterogeneous.