Recent Developments in Cluster-Robust Inference

A. Colin Cameron and Douglas L. Miller .
Univ. of California - Davis, Dept. of Economics
Cornell University, Brooks School of Public Policy and Dept. of Economics

Presented at 2022 Stata Economics Virtual Symposium

.

These slides are for a survey article that is in preparation.
The slides and references are available at
http://cameron.econ.ucdavis.edu/research/papers.html

November 3, 2022

Introduction

- These slides are for a literature survey in preparation
  - so they are lengthy
  - in this talk I will cover some key points.

Cluster error correlation

- Cluster error correlation
  - errors are correlated within cluster (or group)
  - and independent across clusters
    - in the simplest case of one-way clustering.
- Many (most?) microeconometrics studies have clustered errors.
- 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 very 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.

Basic References

- 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 (2022), "Cluster-robust inference: A guide to empirical practice", *Journal of Econometrics*, in-press.
- Recent texts place more emphasis on cluster-robust methods
  - Bruce E. Hansen (2022), *Econometrics*, Princeton University Press.
  - A. Colin Cameron and Pravin K. Trivedi (2022), *Microeconometrics using Stata*, Second edition, Stata Press.

Outline

1. Leading Examples
2. Basics of Cluster-Robust Inference for OLS
3. Better Cluster-Robust Inference for OLS
4. Beyond One-way Clustering
5. Estimators other than OLS
6. Conclusion

1. Example 1: Individuals in Cluster

- Example: How do job injury rates effect wages? Hersch (1998).
  - CPS individual data on male wages.
  - But there is no individual data on job injury rate.
  - Instead aggregated data on occupation injury rates 211
- OLS estimate model for individual i in occupation g

$$y_{ig} = \alpha + \mathbf{x}_{ig}'\boldsymbol{\beta} + \gamma \times z_g + u_{ig}.$$

- Problem:
  - the regressor z_g (job injury risk in occupation g) is perfectly correlated within cluster (occupation)
    - by construction
  - and the error u_{ig} is (mildly) correlated within cluster
    - if model overpredicts for one person in occupation g it is likely to overpredict for others in occupation g.

- Simpler model, nine occupations, $N = 1498$.
- Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
lnw	1498	2.455199	.559654	1.139434	4.382027
occrate	1498	3.208274	2.990179	.461773	10.78546
potexp	1498	19.91288	11.22332	0	53.5
potexpsq	1498	522.4017	516.9058	0	2862.25
educ	1498	12.97296	2.352056	3	20
union	1498	.1321762	.3387954	0	1
nonwhite	1498	.1008011	.3011657	0	1
northe	1498	.2503338	.4333499	0	1
midw	1498	.2683578	.4432528	0	1
west	1498	.2089453	.406691	0	1
occ_id	1498	182.506	99.74337	63	343

- Same OLS regression with different se's estimated using Stata
  - (1) i.i.d. errors, (2) het errors, (3,4) clustered errors

```
global covars potexp potexpsq educ union nonwhite northe midw west
regress lnw occrate $covars
estimates store one_iid
regress lnw occrate $covars, vce(robust)
estimates store one_het
regress lnw occrate $covars, vce(cluster occ_id)
estimates store one_clu
xtset occ_id
xtreg lnw occrate $covars, pa corr(ind) vce(robust)
estimates store one_xtclu
estimates table one_iid one_het one_clu one_xtclu, ///
   b(%10.4f) se(%10.4f) p(%10.3f) stats(N N_clust rank F)
```

- Same OLS coefficients but
  - cluster-robust standard errors (columns 3 and 4) when cluster on occupation are 2-4 times larger than default (column 1) or heteroskedastic-robust (column 2)
  - and some p-values in the last two columns differ substantially: $t(8)$ (column 3) versus $N(0,1)$ (column 4)

Variable	one_iid	one_het	one_clu	one_xtclu
occrate	-0.0448	-0.0448	-0.0448	-0.0448
	0.0044	0.0044	0.0164	0.0163
	0.000	0.000	0.026	0.006
potexp	0.0420	0.0420	0.0420	0.0420
	0.0039	0.0037	0.0073	0.0073
	0.000	0.000	0.000	0.000
potexpsq	-0.0006	-0.0006	-0.0006	-0.0006
	0.0001	0.0001	0.0001	0.0001
	0.000	0.000	0.000	0.000
educ	0.0840	0.0840	0.0840	0.0840
	0.0055	0.0065	0.0175	0.0175
	0.000	0.000	0.001	0.000
union	0.2557	0.2557	0.2557	0.2557
	0.0362	0.0336	0.0892	0.0889
	0.000	0.000	0.021	0.004

- And cluster-robust variance matrix is rank deficient

nonwhite	-0.1057	-0.1057	-0.1057	-0.1057
	0.0391	0.0369	0.0502	0.0501
	0.007	0.004	0.068	0.035
northe	0.0501	0.0501	0.0501	0.0501
	0.0326	0.0340	0.0225	0.0224
	0.125	0.141	0.057	0.025
midw	-0.0124	-0.0124	-0.0124	-0.0124
	0.0319	0.0329	0.0300	0.0299
	0.698	0.707	0.691	0.679
west	0.0402	0.0402	0.0402	0.0402
	0.0339	0.0347	0.0370	0.0369
	0.236	0.246	0.309	0.276
_cons	0.9679	0.9679	0.9679	0.9679
	0.0876	0.1014	0.2461	0.2453
	0.000	0.000	0.004	0.000
N	1498	1498	1498	1498
N_clust			9.0000	
rank	10.0000	10.0000	8.0000	8.0000
F	95.2130	89.0902	.	

legend: b/se/p

- Moulton (1986, 1990) is key paper to highlight the larger standard errors when cluster
  - due to regressors correlated within cluster and errors correlated within cluster.
- The different p-values in columns 3 and 4 arise when there are few clusters
  - use $t(8)$ or more generally $t(G-1)$ not $N(0,1)$
- The rank deficiency of the overall F-test is explained below
  - individual t-statistics are still okay.

Example 2: Difference-in-Differences State-Year Panel ("BDM Setting")

- Example: How do wages respond to a policy indicator variable d_{ts} that varies by state?
  - e.g. $d_{ts} = 1$ if minimum wage law in effect
- OLS estimate model for state s at time t

$$y_{ts} = \alpha + \mathbf{x}'_{ts}\boldsymbol{\beta} + \gamma \times d_{ts} + u_{ts}.$$

- Problem:
  - the regressor d_{ts} is highly correlated within cluster
    - typically d_{ts} is initially 0 and at some stage switches to 1
  - the error u_{ts} is (mildly) correlated within cluster
    - if model underpredicts for California in one year then it is likely to underpredict for other years.

- Again find that default OLS standard errors are way too small
  - should instead do cluster-robust (cluster on state)
- The same problem arises if we have data in individuals (i) in states and years

$$y_{its} = \alpha + \mathbf{x}_{its}'\boldsymbol{\beta} + \gamma \times d_{ts} + u_{its}$$

  - in that case should again cluster on state.
- Bertrand, Duflo & Mullainathan (2004) key paper that highlighted problems for DiD
  - in 2004 people either ignored the problem or with *its* data erroneously clustered on state-year pair and not on state.

2.1 Intuition for cluster-robust inference

- Consider the sample mean $\widehat{\mu}=\bar{y}$ given data $y_i \sim (\mu, \sigma^2)$.

$$\mathrm{Var}[\widehat{\mu}] = \mathrm{Var}[\bar{y}] = \mathrm{Var}\left[\tfrac{1}{N}\textstyle\sum_{i=1}^N y_i\right] = \tfrac{1}{N^2}\left[\textstyle\sum_{i=1}^N \sum_{j=1}^N \mathrm{Cov}(y_i, y_j)\right].$$

- Clustering with equicorrelation ("'exchangeble errors"):
  $\mathrm{Cov}(y_i, y_j) = \rho\sigma^2$ for $i \neq j$

  So $\mathrm{Var}[\mathbf{y}] = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$

$$\begin{aligned} \text{and } \mathrm{Var}[\bar{y}] &= \tfrac{1}{N^2}\left[\textstyle\sum_{i=1}^N \mathrm{Var}(y_i) + \sum_{i=1}^N \sum_{j=1; j\neq i}^N \mathrm{Cov}(y_i, y_j)\right] \\ &= \tfrac{1}{N^2}[N\sigma^2 + N(N-1)\rho\sigma^2] = \tfrac{1}{N}\sigma^2\{1+(N-1)\rho\}. \end{aligned}$$

- $\mathrm{Var}[\bar{y}] > \frac{1}{N}\sigma^2$ and the multiplier grows linearly in N and ρ
  - e.g. $\rho = 0.1$ and $N = 81$ then $\mathrm{Var}[\bar{y}] = 9 \times (\frac{1}{N}\sigma^2)$.

2.2 Cluster-robust variance matrix for OLS

- Linear model for G clusters with N_g individuals per cluster

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

- 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

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

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.

When to Cluster and at what level

- Rule of thumb: with one-way clustering then approximately the incorrect default OLS variance estimate should be inflated by

$$\tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1)$$

  - (1) ρ_{x_j} is the within-cluster correlation of regressor x_j
  - (2) ρ_u is the within-cluster error correlation
  - (3) $\bar{N}_g$ is the average cluster size.
  - Need both (1) and (2) and it increases linearly with (3).
- This result provides very useful guidance in practice!
  - though strictly speaking it is within cluster correlation of $x_j u$ that matters.
- It is not always obvious how to specify the clusters.
  - cluster at the level of an aggregated regressor
  - cluster at the highest level where there may be correlation
    - ★ e.g. for individual in household in state may want to cluster at level of the state if state policy variable is a regressor of interest.

2.3 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
    - ★ e.g. if $u_{ig} = \lambda_{ig}\alpha_g + \varepsilon_{ig}$ rather than simpler $u_{ig} = \alpha_g + \varepsilon_{ig}$.
  - can complicate proofs beyond one-way cluster for OLS.

2.4 Confidence Intervals and Hypothesis Tests

- For a single coefficient β, asymptotic theory gives

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

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

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

- 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.

2.5 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.
- Randomized control trials are often clustered
  - treatment within cluster may be homogeneous or may be heterogeneous.

2.6 Cluster-Specific Fixed Effects Models: Summary

- Now $y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + \alpha_g + u_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + \sum_{h=1}^{G} \alpha_g dh_{ig} + u_{ig}$.
- 1. **FE's do not in practice absorb all within–cluster correlation** of u_{ig}
  - **still need to use cluster-robust VCE.**
- 2. Cluster-robust VCE is still okay with FE's (if $G \to \infty$)
  - Arellano (1987) for N_g small and Hansen (2007a, p.600) for $N_g \to \infty$
- 3. If N_g is small use `xtreg, fe` not `reg i.id_clu` or `areg`
  - as `reg` or `areg` uses wrong degrees of freedom.
- 4. FGLS with fixed effects needs to bias-adjust for $\widehat{\alpha}_g$ inconsistent.
- 5. Need to do a modified Hausman test for fixed effects.
- 6. Modify with `idcluster` option if bootstrapping.
- 7. Several ways of dealing with many two-way fixed effects
  - `reg2hdfe`, `felsdvreg`, McCaffrey et al. (SJ, 2012) review.

3. Better One-way Cluster-Robust Inference

- Consider two-sided symmetric t-test

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

- Three primary challenges to obtaining correct inference
  - $\mathrm{se}(\widehat{\beta})$ has many-cluster bias
  - $\mathrm{se}(\widehat{\beta})$ has few-cluster bias
  - $\mathrm{se}(\widehat{\beta})$ 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: Many-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.

3.2 Challenge 2: Few-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).

CR3 Standard Errors

- The $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}(\sum_{g=1}^{G} \mathbf{X}_g' \widetilde{\mathbf{u}}_g \widetilde{\mathbf{u}}_g' \mathbf{X}_g)(\mathbf{X}'\mathbf{X})^{-1}$.
- Bell and McCaffrey (2002) instead use

$$\widetilde{\mathbf{u}}_g = \sqrt{\frac{G-1}{G}}[\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g .$$

- Then $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}]$ is equivlent to the jackknife estimate of the variance of the OLS estimator
  - where $\widehat{\boldsymbol{\beta}}^{(g)}$ are delete-one-cluster estimates of $\boldsymbol{\beta}$

$$\widehat{\mathrm{V}}_{\mathrm{CR3}}[\widehat{\boldsymbol{\beta}}] = \frac{G-1}{G} \sum_{g=1}^{G} (\widehat{\boldsymbol{\beta}}^{(g)} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^{(g)} - \widehat{\boldsymbol{\beta}})'$$

- Recent research finds that this works well
  - MacKinnon, Nielsen and Webb (2022, *JE*)
  - Hansen (2022, WP) proves that CR3 is never downward biased whereas CR1 can be extremely downward biased.
- In Stata: `vce(jackknife,mse)`
- Fast implementation: MacKinnon, Nielsen and Webb (2022, QED WP 1485)

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 (2022, JE, 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.
- Sasaki and Wang (2022, WP) find that a small number of large clusters leads to violation of the moment assumptions used to prove consistency of standard CRVE of OLS and instead proposed weighted LS estimator.

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 $N_g \longrightarrow \infty$
  - 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 (or "studentized") 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 $\widehat{\beta}$ 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 $\widehat{\mathbf{y}}^*_g = \mathbf{X}'_g \widehat{\beta} + \widehat{\mathbf{u}}^*_g$.
      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 $N_g \rightarrow \infty$ 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 $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + \mathbf{z}_g'\gamma + u_{ig}$.
  - 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 that 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.
- Young (2022, WP) considers interactions between treatment effects and covariates.
- Toulis (2022, WP) uses randomization with exchangeable errors within cluster.

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.

3.7 Design-based inference

- AAIW (2022, *QJE*) discussed below propose alternative inference methods that can lead to substantially smaller cluster-robust standard errors than traditional inference.
- Let $Y = f(U, Z, \varepsilon)$ where
  - U is treatment variable
  - Z is other variables (called "attributes" rather than "controls")
  - ε is error.
- Randomness may potentially come from U, Z, ε and from sample S from the population.
- Traditional approaches
  - randomness is due to model errors ε (called "model" approach)
  - randomness is due to selection of sample S from the population
    - ★ problem if sample is the population e.g. states.
- Design-based approach (newer)
  - randomness is due to assignment of treatment U.

Pure design-based inference

- Suppose randomness comes solely from treatment assignment.
- Neyman (1923, English translation 1990) had two innovations
  - a potential outcomes framework (though did not call it that)
  - designed-based inference that treats potential outcomes as nonrandom
    - so not "model-based" with a model error term
    - instead randomness comes solely from treatment assignment.
- For binary treatment
  - $\text{Var}[\bar{y}_1 - \bar{y}_0] = \text{Var}[y_{1i}]/n_1 + \text{Var}[y_{0i}]/n_0 - \text{Var}[y_{1i} - y_{0i}]/(n_0 + n_1)$
  - is less than standard $\text{Var}[y_{1i}]/n_1 + \text{Var}[y_{0i}]/n_0$ if there is heterogeneous treatment effect
  - though $\text{Var}[y_{1i} - y_{0i}]$ is inestimable (without further assumptions)
  - Imbens and Rubin (2015, ch.6) detail this.

Design-based inference plus sampling-based inference

- Abadie, Athey, Imbens, Wooldridge (2020, Ecta)
  - independent observations as for Neyman (1923)
  - design-based treatment and no model error as for Neyman (1923)
  - **add sampling-based inference**
    - allows for a subset of a finite population to be sampled
    - Neyman instead implicitly viewed entire population as sampled.
- They obtain a variance estimate $V_{causal,sample}[\widehat{\theta}]$ that
  - is generally less than Eicker-Huber-White $V_{EHW}[\widehat{\theta}]$
  - is nonzero even if the entire population is sampled
    - because across repeated samples the treatment varies, leading to different potential outcomes being chosen
  - equals $V_{EHW}[\widehat{\theta}]$ if sample treatment effects are constant
  - equals $V_{EHW}[\widehat{\theta}]$ if the fraction sampled goes to zero
  - is approximately 65% of $V_{EHW}[\widehat{\theta}]$ in AAIW's simulations.

Detail for AAIW (2020)

- $Y_i^*(\cdot)$ are potential outcomes, U_i is treatment, $Y_i = Y_i^*(U_i)$ is observed.
- Introduce "attributes" Z_i (includes intercept)
  - these are needed to provide an estimate of B_{cond} given below.
- Define $X_i = U_i - \widehat{U}_i$ where $\widehat{U}_i$ prediction from regress $E[U_i]$ on $\mathbf{Z}_i$.
- OLS of Y_i on X_i and Z_i gives same θ as OLS of Y_i on U_i and $\mathbf{Z}_i$.
- Define residual $\varepsilon_i = Y_i - \theta X_i - \mathbf{Z}_i' \gamma$.
- Theory views a sequence of samples each drawn from the same population with n fixed observations on $\mathbf{Y}$, $\mathbf{U}$, $\mathbf{Z}$.
- $\mathrm{V}_{EHW}[\widehat{\theta}] = A^{-1} B_{EHW} A^{-1}$ where $B_{EHW} = \lim \frac{1}{n} \sum_{i=1}^n E[\varepsilon_i^2 X_i^2]$
- $\mathrm{V}_{causal,sample}[\widehat{\theta}] = A^{-1} B_{cond} A^{-1}$ where $B_{cond} = \lim \frac{1}{n} \sum_{i=1}^n Var[X_i \varepsilon_i]$
- $B_{EHW} - B_{cond} = \lim \frac{1}{n} \sum_{i=1}^n E[\varepsilon_i X_i] \times E[\varepsilon_i X_i]$ is pos. semidefinite.
- In practice can only conservatively estimate B_{cond}.

Additionally allow model error

- Starz and Steigerwald (2022, WP)
  - independent observations
  - extend AAIW (2020, Ecta) by **bringing in possible model error.**
- Let θ be the average treatment effect (ATE) in the population.
- Then the variance of $\widehat{\theta}$ has two components
  - AAIW-like term that controls for treatment assignment and finite sampling
    $+$ standard OLS result due to model error.
- The estimate of the variance of $\widehat{\theta}$ then varies with the proportion of shocks due to the model error
  - if all is due to model errors then use the usual robust VCE
  - if none is due to model errors then $\widehat{\mathrm{V}}[\widehat{\theta}]$ can be much smaller, especially if there is considerable heterogeneity and/or most of the population is sampled.

Details for Starz and Steigerwald (2022)

- Consider simplest case of the sample mean (so no treatment)
  - sampling binary indicator R_i is Bernoulli with $\rho = \Pr[R_i = 1]$
  - random error so $Y_i = y_i + \varepsilon_i$ where $E[y_i] = \mu$ and ε_i is i.i.d. $(0, \sigma_\varepsilon^2)$.
- Estimator of μ is $\widehat{\mu}_n = (\frac{1}{n}\sum_{i=1}^n R_i Y_i) / \ (\frac{1}{n}\sum_{i=1}^n R_i)$.
- Then $\mathrm{Var}[\widehat{\mu}_n] = (1-\rho)\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2/\rho n \ + \ \sigma_\varepsilon^2/\rho n$
  - first term is usual finite sampling term
    - goes to zero if $\rho \to 1$ or heterogeneity in $y \to 0$
  - second term is usual formula for variance of the mean
- First term is estimated by $(1-\frac{N}{n})\frac{\widehat{s}^2}{N}$ where $\widehat{s}^2 = \frac{1}{N}\sum_{i=1}^n R_i(Y_i - \widehat{\mu}_n)^2$
  - this gives lower bound for $\widehat{\mathrm{V}}[\widehat{\mu}_n]$ of $(1-\frac{N}{n})\frac{\widehat{s}^2}{N}$ if $\sigma_\varepsilon^2 = 0$.
- The second term is estimated by $\frac{N}{n}\frac{\widehat{s}^2}{N}$
  - this gives upper bound for $\widehat{\mathrm{V}}[\widehat{\mu}_n]$ of $(1-\frac{N}{n})\frac{\widehat{s}^2}{N} + \frac{N}{n}\frac{\widehat{s}^2}{N} = \frac{\widehat{s}^2}{N}$.

Clustered data and design-based plus sampling-based inference

- Abadie, Athey, Imbens, Wooldridge (2022, *QJE*, revision of 2017, NBER WP) "When Should You Adjust Standard Errors for Clustering".
    - extends AAIW(2020) by considering the clustered case.
- Estimate the population average treatment effect θ using $\widehat{\theta} = \bar{y}_1 - \bar{y}_0$
- Define $\text{Var}[\widehat{\theta}]$ to be the limiting variance under the assumptions
    - sampling: sample clusters and then sample units within chosen clusters
    - treatment: binary treatment may be correlated within cluster
    - model error: none.
- Then $\widehat{\text{V}}_{\text{CR}}[\widehat{\theta}]$ (the usual cluster-robust VCE) **can greatly over-estimate** $\text{Var}[\widehat{\theta}]$
    - though not if only a few clusters in the population are sampled
    - and not if treatment effects are constant across clusters
    - and not if all units in a cluster receive the same treatment.

Details on AAIW (2022)

- Potential outcomes with binary treatment
  - $Y_i^*(\cdot)$ are potential outcomes (2022 paper uses $y_i(\cdot)$)
  - $U_i = (0,1)$ is stochastic binary treatment (2022 paper uses $W_i(\cdot)$)
  - $Y_i = Y_i^*(U_i)$ is observed
  - i denotes individual unit and m denotes cluster.
- Sampling process
  - $R_i = (0,1)$ is stochastic sample inclusion
    - first sample cluster with probability $q \in (0,1]$
    - second sample units in chosen clusters with probability $p \in (0,1]$.
- Treatment assignment process
  - $U_i = (0,1)$ is set to one with with random probability $A_m \in [0,1]$
  - the cluster-specific probability A_m is drawn from (μ, σ^2) distribution
    - assignment is correlated within cluster if $\sigma^2 > 0$.

Details on AAIW (2022) continued

- Let $\widehat{\mathrm{V}}_{\mathrm{CCV}}[\widehat{\theta}]$ denote the newly proposed estimate.
- When $q = 1$ do the following two-step bootstrap resampling procedure.
- At replication b
  - 1. For each cluster $m = 1, ..., M$ draw the cluster-level fraction treated $\overline{U}_m^b$ with replacement from the sample cluster-level fractions $\overline{U}_1^b, ..., \overline{U}_M^b$.
  - 2. For each cluster $m = 1, ..., M$ with $\overline{N}_m$ units draw with replacement $\overline{N}_m \overline{U}_m^b$ units from the treated and $\overline{N}_m(1 - \overline{U}_m^b)$ from the untreated.
- When $q < 1$ (so not all clusters in population are sampled)
  - adapt the above as given in paper section 4.3
  - use a linear combination of the new $\widehat{\mathrm{V}}_{\mathrm{CCV}}[\widehat{\theta}]$ and the usual $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\theta}]$ with weights q and $(1 - q)$.

Comments on AAIW (2022)

- The method can make a big difference when most clusters are sampled, treatment varies within cluster, treatment effects vary across clusters and there are many observations per cluster.
- U.S. cross-section example with all 52 states, 50,000 observations average per state, binary treatment at individual level and not state level
  - usual cluster-robust se is 7 times larger than new CCV se
  - and with state fixed effects usual cluster-robust se is 20 times CCV.
- Main critiques would be conceptual
  - is there no role for a model error?
  - the new method assumes that the probability of an individual in California receiving treatment is a random draw from the empirical distribution of the treatment fractions for the 52 states.
- And generalizability
  - e.g. to panel data (static and dynamic).

Design-based approach with cluster-level treatment assignment

- Su and Ding (2021, JRSSB)
  - designed-based inference (no model error and no sampling issues)
  - treatment assignment: units in a cluster are either all treated or all not treated.
- Consider the efficiency of various estimators of the ATE
  - should estimators be at individual level or use cluster averages (possibly weighted)
  - add control variables ("model-assisted") to improve efficiency
    - these are unnecessary for consistent estimation as we consider an RCT.
- Favors regression based on cluster totals.

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 high-dimensional 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 (2022, WP)
  - 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)).
- Davezies, D'Haultfoeuille and Guyonvarch (2021, *AS*)
  - provides empirical process theory that assumes exchangeability and propose a pigeonhole bootstrap.
- Chiang, Kato and Sasaki (2021, *JASA*)
  - inference and bootstraps for high-dimensional exchangeable arrays.
- 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

- 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.
- Finally, consider machine learning.

5.1 Feasible GLS

- Potential efficiency gains for feasible GLS compared to OLS.
- And for one-way clustering there is a cluster-robust VCE (as $G \rightarrow \infty$)

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

- 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.

5.6 Machine learning prediction and clustering

- Cameron and Trivedi (2022, chapter 28) provide an accessible introduction to machine learning.
- Leading ML methods used by econometricians in order of current usage
  - lasso (and to a lesser extent ridge)
  - random forests (collections of regression trees)
  - neural networks (including deep nets).
- For lasso linear regression with independent data choose $\boldsymbol{\beta}$ to minimize
  - $Q_\lambda(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\kappa_j|\beta_j|$
    - where in the simplest case the regressors are standardized and $\kappa_j = 1$.
- With clustered data we could use the same objective function.
- Stata instead uses a weighted average
  - $Q_\lambda(\beta) = \frac{1}{G}\sum_{g=1}^{G}\left\{\frac{1}{N_g}\sum_{i=1}^{N_g}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right\} + \lambda\sum_{j=1}^{p}\kappa_j|\beta_j|$
  - same as simple unweighted in the case of balanced clusters.

Causal machine learning

- A key general paper for double/debiased ML is Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, J. (2018, EJ).
- A leading example is the partial linear model with scalar regressor of interest d and many potential controls $\mathbf{x}_c$
  - $y = \alpha d_i + g(\mathbf{x}_c) + u$ where $g(\cdot)$ is unspecified.
- Then
  - a machine learner is used to approximate $g(\mathbf{x}_c)$
  - estimation of α is based on an "orthogonalized" moment condition that enables standard inference on α despite the first-stage use of a a machine learner
  - performance is improved by using cross fitting
    - ★ a bigger part of the data is used in the ML stage and the smaller remainder is used in second stage estimation of α.

Causal machine learning and clustered data

- With clustering the cross fitting needs to be adapted.
- For one-way clustering (such as panel data)
    - Belloni, Chernozhukov, Hansen, and Damien Kozbur (2016, *JBES*)
    - cross fitting keeps clusters intact.
- For two-way clustering (such as panel data)
    - Chiang, Kato, Ma and Sasaki (2022, *JBES*)
    - cross fitting in simplest case splits sample in each direction in half giving $2^2 = 4$ distinct groups.
- For dyadic clustering (such as panel data)
    - Chiang, Kato, Ma and Sasaki (2022, WP)
    - a more complex cross fitting is proposed.
- Recent work challenges sparsity assumption and develops alternative inference for regular OLS
    - Cattaneo, Jansson and Newey (2018b, *JASA*), Li and Müller (2021a, *QE*), Riccardo D'Adamo (2019, WP).

6. Conclusion

- Where clustering is present it is important to control for it.
- Most empirical 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.