Robust Inference with Clustered Errors

Colin Cameron
Univ. of California - Davis

Keynote address at The 8th Annual Health Econometrics Workshop,
University of Colorado Denver, Anschutz Medical Campus.

.

Based on A Practitioners Guide to Cluster-Robust Inference
J. of Human Resources, 2015, vol.50, 317-372.
Joint work with Douglas L. Miller (and earlier Jonah Gelbach).

September 30 2016

1. Introduction

- Consider straightforward OLS estimation in linear regression model.
- Suppose estimator $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.
- Concerned with getting the correct standard errors of $\widehat{\boldsymbol{\beta}}$
  - default: if errors are i.i.d. $(0, \sigma^2)$
  - heteroskedastic-robust: if errors are independent $(0, \sigma_i^2)$
  - heteroskedastic and autocorrelation-robust (HAC): if errors are serially correlated
  - cluster-robust: if errors are correlated within cluster and independent across clusters
    - ★ this talk.

- Why is this important?
- 1. Cluster-robust standard errors can be much bigger than default or heteroskedastic-robust.
- 2. So failure to control for clustering
  - overstates t statistics and understates p-values
  - provides too narrow confidence intervals
- 3. This arises often especially in the empirical / public labor literature using quasi-experimental methods.
- 4. There are subtleties - not always straightforward to implement.

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 j it is likely to overpredict for others in occupation j.

- 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) iid 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 p-values in the last two columns differ substantially: $t(8)$ versus $N(0,1)$

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 $t(8)$ 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

- 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 also 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 state.

Outline

2. Cluster-Robust Inference for OLS

- Clustered errors: $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + u_{ig}$ with u_{ig} correlated with error for any observation in group g and uncorrelated with error for any observation in other groups.
- Key result is that then 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 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 also increases with (3).
- Cluster-robust estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ is natural extension of White's (1980) heteroskedastic-robust estimate
  - but requires number of groups $G \longrightarrow \infty$.
- Potentially more efficient feasible GLS is possible and can also be robustified.

2.1 Intuition

- Suppose we have univariate data $y_i \sim (\mu, \sigma^2)$.
- We estimate μ by $\bar{y}$ and

$$\text{Var}[\bar{y}] = \text{Var}\left[\frac{1}{N}\sum_{i=1}^{N} y_i\right] = \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\text{Cov}(y_i, y_j)\right].$$

- Given independence over i this simplifies to $\text{Var}[\bar{y}] = \frac{1}{N}\sigma^2$.
- Now suppose observations are equicorrelated with $\text{Cov}(y_i, y_j) = \rho\sigma^2$

for $i \neq j$ so $\text{Var}[\mathbf{y}] = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$. Then

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

- So independent errors
$$\text{Var}[\bar{y}] = \frac{1}{N}\sigma^2.$$
- Equicorrelated errors
$$\text{Var}[\bar{y}] = \frac{1}{N}\sigma^2\{1 + (N-1)\rho\}.$$
- The variance is $1 + (N-1)\rho$ times larger!.
- Reason: An extra observation is not providing a new independent piece of information.
- Note that the effect can be large
  - if $\rho = 0.1$ (so R^2 of y_i on y_j is 0.01)
  - and $N = 81$
  - then $\text{Var}[\bar{y}] = 9 \times \frac{1}{N}\sigma^2$ is 9 times larger!

2.2 OLS with Clustered Errors

- Model for G clusters with N_g individuals per cluster:

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

- OLS estimator

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\textstyle\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{x}_{ig}\mathbf{x}_{ig}')^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{x}_{ig}y_{ig}) \\ &= (\textstyle\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g)^{-1}(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

- As usual

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}(\textstyle\sum_{g=1}^{G}\mathbf{X}_g\mathbf{u}_g).\end{aligned}$$

- Assume independence over g and correlation within g

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

- Then $\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, \mathsf{V}[\widehat{\boldsymbol{\beta}}]]$ with asymptotic variance

$$\begin{aligned}\text{Avar}[\widehat{\boldsymbol{\beta}}] &= (\mathsf{E}[\mathbf{X}'\mathbf{X}])^{-1}(\textstyle\sum_{g=1}^{G}\mathsf{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g])(\mathsf{E}[\mathbf{X}'\mathbf{X}])^{-1} \\ &\neq \sigma_u^2(\mathsf{E}[\mathbf{X}'\mathbf{X}])^{-1}.\end{aligned}$$

Consequences - KEY RESULT FOR INSIGHT

- Suppose equicorrelation within cluster g

$$\text{Cor}[u_{ig}, u_{jg}|\mathbf{x}_{ig}, \mathbf{x}_{jg}] = \begin{cases} 1 & i = j \\ \rho_u & i \neq j \end{cases}$$

  - this arises in a random effects model with $u_{ig} = \alpha_g + \varepsilon_{ig}$, where α_g and ε_{ig} are i.i.d. errors.
  - an example is individual i in village g or student i in school g.

- 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 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 also increases with (3)

- Theory: Kloek (1981), Scott and Holt (1982).
- Practice: Moulton (1986, 1990) showed that the variance inflation can be large even if ρ_u is small
  - especially with a grouped regressor (same for all individuals in group) so that $\rho_x = 1$.
  - CPS data example:
    $N_g = 81$, $\rho_x = 1$ and $\rho_u = 0.1$
    $\Longrightarrow \tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1) = 1 + 1 \times 0.1 \times 80 = 9.$
    - ★ true standard errors are three times the default!
- So should correct for clustering even in settings where not obviously a problem.

2.3 The Cluster-Robust Variance Matrix Estimate

- Recall for OLS with independent heteroskedastic errors

$$\text{Avar}[\widehat{\boldsymbol{\beta}}] = (\text{E}[\mathbf{X}'\mathbf{X}])^{-1}(\textstyle\sum_{i=1}^{N} \text{E}[u_i^2\mathbf{x}_i\mathbf{x}_i'])(\text{E}[\mathbf{X}'\mathbf{X}])^{-1}$$

  can be consistently estimated (White (1980)) as $N \longrightarrow \infty$ by

$$\widehat{\text{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}(\textstyle\sum_{i=1}^{N} \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i')(\mathbf{X}'\mathbf{X})^{-1}.$$

- Need $\frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i' - \frac{1}{N}\sum_{i=1}^{N}\text{E}[u_i^2\mathbf{x}_i\mathbf{x}_i'] \xrightarrow{p} 0$
  - not $\widehat{u}_i^2 \xrightarrow{p} \text{E}[u_i^2]$

- Similarly for OLS with independent clustered errors

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

can be consistently estimated as $G \longrightarrow \infty$ by the cluster-robust variance estimate (CRVE)

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

  - Stata uses $\widetilde{\mathbf{u}}_g = c\widehat{\mathbf{u}}_g = c(\mathbf{y}_g - \mathbf{X}_g\widehat{\boldsymbol{\beta}})$ where $c = \frac{G}{G-1}\frac{N-1}{N-K} \simeq \frac{G}{G-1}$.

- The CRVE was
  - proposed by White (1984) for balanced case
  - proposed by Liang and Zeger (1986) for grouped data
  - proposed by Arellano (1987) for FE estimator for short panels (group on individual)
  - Hansen (2007a) and Carter, Schnepel and Steigerwald (2013) also allow $N_g \longrightarrow \infty$.
  - popularized by incorporation in Stata as the cluster option (Rogers (1993)).
  - also allows for heteroskedasticity so is cluster- and heteroskedastic-robust.
- Stata with cluster identifier `id_clu`
  - `regress y x, vce(cluster id_clu)`
  - `xtreg y x, pa corr(ind) vce(robust)`
    - ★ after `xtset id_clu`
    - ★ from version 12.1 on Stata interprets `vce(robust)` as cluster-robust for all `xt` commands.

2.4. Feasible GLS with Cluster-Robust Inference

- Potential efficiency gains for feasible GLS compared to OLS.
- Specify a model for $\Omega_g = \mathrm{E}[\mathbf{u}_g\mathbf{u}_g'|\mathbf{X}_g]$, e.g. within-cluster equicorrelation.
- Given $\widehat{\Omega} \xrightarrow{p} \Omega$, the feasible GLS estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{FGLS}} = \left(\sum_{g=1}^{G} \mathbf{X}_g'\widehat{\Omega}_g^{-1}\mathbf{X}_g\right)^{-1} \sum_{g=1}^{G} \mathbf{X}_g'\widehat{\Omega}_g^{-1}\mathbf{y}_g.$$

- Default $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{\text{FGLS}}] = (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}$ requires correct Ω.
- To guard against misspecified Ω_g use cluster-robust

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

  - where $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\widehat{\boldsymbol{\beta}}_{\text{FGLS}}$ and $\widehat{\Omega} = \text{Diag}[\widehat{\Omega}_g]$
  - assumes $\mathbf{u}_g$ and $\mathbf{u}_h$ are uncorrelated, for $g \neq h$
  - and needs $G \to \infty$.

FGLS Examples

- Example 1 - Moulton setting
  - Random effects model: $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g + \varepsilon_{ig}$
    - ★ `xtreg, re vce(robust)`
  - Richer hierarchical linear model or mixed model
    - ★ Stata 13: `mixed, vce(robust)`
- Example 2 - BDM setting
  - AR(1) error $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$ and ε_{it} i.i.d.
  - `xtreg y x, pa corr(ar 1) vce(robust)`
  - Stata allows a range of correlation structures
- Puzzle - why is FGLS not used more?
  - Easily done in Stata with robust VCE if $G \longrightarrow \infty$
  - Unless FE's present and N_g small (see later).

2.5 The CRVE can be rank deficient

- $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}]$ can be rank deficient
  - rank is as most minimum of K and $G-1$
  - $\widehat{\mathbf{B}} = \mathbf{C}'\mathbf{C}$, where $\mathbf{C}' = [\mathbf{X}_1'\widehat{\mathbf{u}}_1 \cdots \mathbf{X}_G'\widehat{\mathbf{u}}_G]$ is $K \times G$
  - and $\mathbf{X}_1'\widehat{\mathbf{u}}_1 + \cdots + \mathbf{X}_G'\widehat{\mathbf{u}}_G = \mathbf{0}$
- For example if have 15 clusters (say states)
  - Cannot jointly test significance of 20 occupation dummies
  - But can test joint significance of 14.
- The test of overall joint statistical significance is not computable if $G < K$
  - but tests on individual coefficients are still okay.

2.6 Pairs Cluster Bootstrap

- Do the following steps for each of B bootstrap samples:
  - (1) form G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters
  - (2) compute $\widehat{\boldsymbol{\beta}}_b$ (estimate of $\boldsymbol{\beta}$) in the b^{th} bootstrap sample.
- Compute the variance of the B estimates $\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_B$ as

$$\widehat{\mathrm{V}}_{\mathrm{CR;boot}}[\widehat{\boldsymbol{\beta}}] = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\beta}}_b - \overline{\widehat{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}_b - \overline{\widehat{\boldsymbol{\beta}}})',$$

  where $\overline{\widehat{\boldsymbol{\beta}}} = B^{-1}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}_b$ and $B \geq 400$.
- Pairs cluster bootstrap has no asymptotic refinement.
  - But can compute these if Stata doesn't provide a CRVE.
  - Also can do even if usual CRVE is rank deficient?
- Also cluster jackknife.

3. 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 uce cluster-robust VCE
- 2. Cluster-robust VCE is still okay with FE's (if $G \longrightarrow \infty$)
  - Arellano (1987) for N_g small and Hansen (2007a, p.600) for $N_g \longrightarrow \infty$
- 3. If N_g small use `xtreg, fe` not `reg i.id_clu`
  - as `reg` or `areg` uses wrong degrees of freedom
- 4. FGLS with fixed effects needs to bias-adjust for $\widehat{\alpha}_g$ inconsistent
  - Hansen (2007b) provides bias-corrected FGLS for AR(p) errors
  - Brewer, Crossley and Joyce (2013) implement in DiD setting
  - Hausman and Kuersteiner (2008) provide bias-corrected FGLS for Kiefer (1980) error model
- 5. Need to do a modified Hausman test for fixed effects.

4.1 Factors Determining What to Cluster Over

- It is not always obvious how to specify the clusters.
- Moulton (1986, 1990)
  - cluster at the level of an aggregated regressor.
- Bertrand, Duflo and Mullainathan (2004)
  - with state-year data cluster on states (assumed to be independent) rather than state-year pairs.
- Pepper (2002)
  - 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.

4.2 Clustering Due to Survey Design

- Clustering routinely arises with complex survey data.
- Then the loss of efficiency due to clustering is called the design effect
  - This is the inverse of the variance inflation factor given earlier
  - Long literature going back to 1960's
  - CRVE is called the linearization formula
  - Shah, Holt and Folsom (1977) is early reference.
- Complex survey data are weighted
  - often ignore assuming conditioning on **x** handles weighting
- And stratified
  - this improves estimator efficiency somewhat
- Bhattacharya (2005) gives a general GMM treatment.

- Econometricians reasonably
  - 1. Cluster on PSU or higher
  - 2. Sometimes weight and sometimes not
  - 3. Ignore stratification (with slight loss in efficiency)
- Survey software controls for all three.
  - Stata `svy` commands
- Econometricians use regular commands with `vce(cluster)` and possibly `[pweight=1/prob]`

5. Multi-way Clustering

- Example: How do job injury rates effect wages? Hersch (1998).
  - CPS individual data on male wages $N = 5960$.
  - But there is no individual data on job injury rate.
  - Instead aggregated data:
    - ★ data on industry injury rates for 211 industries
    - ★ data on occupation injury rates for 387 occupations.
- Model estimated is

$$y_{igh} = \alpha + \mathbf{x}'_{igh}\boldsymbol{\beta} + \gamma \times rind_{ig} + \delta \times rocc_{ih} + u_{igh}.$$

- What should we do?
  - Ad hoc robust: OLS and robust cluster on industry for $\widehat{\gamma}$ and robust cluster on occupation for $\widehat{\delta}$.
  - Non-robust: FGLS two-way random effects: $u_{igh} = \varepsilon_g + \varepsilon_h + \varepsilon_{igh}$; ε_g, ε_h, ε_{igh} i.i.d.
  - Two-way robust: next

5.1 Two-way Cluster-Robust

- Robust variance matrix estimates are of the form

$$\widehat{\text{Avar}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{B}}(\mathbf{X}'\mathbf{X})^{-1}$$

- For one-way clustering with clusters $g = 1, ..., G$ we can write

$$\widehat{\mathbf{B}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j\mathbf{1}[i, j \text{ in same cluster } g]$$

  - where $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ and
  - the indicator function $\mathbf{1}[A]$ equals 1 if event A occurs and 0 otherwise.

- For two-way clustering with clusters $g = 1, ..., G$ and $h = 1, ..., H$

$$\begin{aligned}\widehat{\mathbf{B}} &= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j\mathbf{1}[i, j \text{ share any of the two clusters}] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j\mathbf{1}[i, j \text{ in same cluster } g] \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j\mathbf{1}[i, j \text{ in same cluster } h] \\ &\quad - \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j\mathbf{1}[i, j \text{ in both cluster } g \text{ and } h].\end{aligned}$$

- Obtain three different cluster-robust "variance" matrices for the estimator by
  - one-way clustering in, respectively, the first dimension, the second dimension, and by the intersection of the first and second dimensions
  - add the first two variance matrices and, to account for double-counting, subtract the third.
  - Thus
$$\widehat{\mathrm{V}}_{\text{two-way}}[\widehat{\boldsymbol{\beta}}] = \widehat{\mathrm{V}}_{G}[\widehat{\boldsymbol{\beta}}] + \widehat{\mathrm{V}}_{H}[\widehat{\boldsymbol{\beta}}] - \widehat{\mathrm{V}}_{G\cap H}[\widehat{\boldsymbol{\beta}}],$$
- Theory presented in Cameron, Gelbach, and Miller (2006, 2011), Miglioretti and Heagerty (2006), and Thompson (2006, 2011)
  - Extends to multi-way clustering.
- Early empirical applications that independently proposed this method include Acemoglu and Pischke (2003).

5.2 Implementation

- If $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ is not positive-definite (small G, H) then
  - Decompose $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = U\Lambda U'$; U contains eigenvectors of $\widehat{\mathrm{V}}$, and $\Lambda = \text{Diag}[\lambda_1, ..., \lambda_d]$ contains eigenvalues.
  - Create $\Lambda^+ = \text{Diag}[\lambda_1^+, ..., \lambda_d^+]$, with $\lambda_j^+ = \max\left(0, \lambda_j\right)$, and use $\widehat{\mathrm{V}}^+[\widehat{\boldsymbol{\beta}}] = U\Lambda^+ U'$
  - Stata add-on `cgmreg.ado` implements this.
- Also Stata add-on `xtivreg2.ado` has two-way clustering for a variety of linear model estimators.
- Fixed effects in one or both dimensions
  - Theory has not formally addressed this complication
  - Intuitively if $G \longrightarrow \infty$ and $H \longrightarrow \infty$ then each fixed effect is estimated using many observations.
  - In practice the main consequence of including fixed effects is a reduction in within-cluster correlation of errors.

Application

- Example 1: Hersch data
  - Relatively small difference versus one-way
  - But can simultaneously handle both ways rather than one-way cluster on industry for $\widehat{\gamma}$ and one-way cluster on occupation for $\widehat{\delta}$.
- Example 2: DiD
  - We have found little difference if cluster two-way on state and time versus just one-way on state.
  - Studies in finance view this as important.
- Example 3: Country-pair international trade volume
  - Two-way cluster on country 1 and country 2 leads to much bigger standard errors (Cameron et al. 2011)
  - Cameron and Miller (2012) find that two-way still doesn't pick up all correlations.
  - Instead other methods including Fafchamps and Gubert (2007).

5.3 Feasible GLS

- Two-way random effects
  - $y_{igh} = \mathbf{x}'_{igh}\boldsymbol{\beta} + \alpha_g + \delta_h + \varepsilon_{ig}$ with i.i.d. errors
  - `xtmixed y x || _all: R.id1 || id2: , mle`.
  - but cannot then get cluster-robust variance matrix
- Hierarchical linear models or mixed models
  - richer FGLS
  - $y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta}_g + u_{ig}$
  - $\boldsymbol{\beta}_g = \mathbf{W}_g\gamma + \mathbf{v}_j$ where u_{ig} and $\mathbf{v}_g$ are errors.
  - see Rabe-Hesketh and Skrondal (2012)

5.4 Spatial Correlation

- Two-way cluster robust related to time-series and spatial HAC.
- In general $\widehat{\mathbf{B}}$ in preceding has the form $\sum_i \sum_j w(i,j)\,\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j$.
  - Two-way clustering: $w(i,j) = 1$ for observations that share a cluster.
  - White and Domowitz (1984) time series: $w(i,j) = 1$ for observations "close" in time to one another.
  - Conley (1999) spatial: $w(i,j)$ decays to 0 as the distance between observations grows.
- The difference: White & Domowitz and Conley use mixing conditions to ensure decay of dependence in time or distance.
  - Mixing conditions do not apply to clustering due to common shocks.
  - Instead two-way robust requires independence across clusters.

Spatial Correlation Consistent VE

- Driscoll and Kraay (1998) panel data when $T \longrightarrow \infty$
  - generalizes HAC to spatial correlation
  - errors potentially correlated across individuals
  - correlation across individuals disappears for obs $> m$ time periods apart
  - then $w(it, js) = 1 - d(it, js)/(m+1)$ with sum over i, j, s and t
  - and $d(it, js) = |t - s|$ if $|t - s| \leq m$ and $d(it, js) = 0$ otherwise.
  - Stata add-on command `xtscc`, due to Hoechle (2007).
- Foote (2007) contrasts various variance matrix estimators in a macroeconomics example.
- Petersen (2009) contrasts methods for panel data on financial firms.
- Barrios, Diamond, Imbens, and Kolesár (2012) state-year panel on individuals with spatial correlation across states. And use randomization inference.

6. Inference with Few Clusters

- One-way clustering, and focus on the Wald "t-statistic"

$$w = \frac{\widehat{\beta} - \beta_0}{s_{\widehat{\beta}}}.$$

- CRVE assumes $G \to \infty$. What if G is small?
- At a minimum use CRVE with rescaled error $\widetilde{\mathbf{u}}_g = \sqrt{c}\widehat{\mathbf{u}}_g$
  - where $c = \frac{G}{G-1}$ or $c = \frac{G}{G-1} \times \frac{N-1}{N-k} \simeq \frac{G}{G-1}$
- And use $T(G-1)$ critical values
  - Stata does this for `regress` but not other commands..
- But tests still over-reject with small G.

- Inference with few clusters
  - arises often in practice e.g. have only ten states
  - standard methods in e.g. Stata over-reject
  - this is an active area of research.
- Three approaches
  - 1. Finite sample bias correction to the CRVE
  - 2. Wild cluster bootstrap (with asymptotic refinement)
  - 3. Better t critical values
- A related distinct problem is one treated cluster and many control clusters.

6.1 The Basic Problem with Few Clusters

- OLS overfits with $\widehat{u}$ systematically biased to zero compared to u.
  - e.g. OLS with iid normal errors $\mathrm{E}[\widehat{\mathbf{u}}'\widehat{\mathbf{u}}] = (N-K)\sigma^2$, not $N\sigma^2$.
- Problem is greatest as G gets small - "few" clusters.
- How few is few?
  - balanced data; $G < 20$ to $G < 50$ depending on data
  - unbalanced data: G less than this.
- Unusual case. If N is too small with cross-section data, usually everything is statistically insignificant.
- With clustered data if G is small we may still have statistical significance if N_g is small.

6.2 Solution 1: Bias-Corrected CRVE

- Simplest is $\widetilde{\mathbf{u}}_g = \sqrt{c}\widehat{\mathbf{u}}_g$, already mentioned.
- CR2VE generalizes HC2 for heteroskedasticity
  - $\widetilde{\mathbf{u}}_g^* = [\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1/2}\widehat{\mathbf{u}}_g$ where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$
  - gives unbiased CRVE if errors iid normal
- CR3VE generalizes HC3 for heteroskedasticity
  - $\widetilde{\mathbf{u}}_g^+ = \sqrt{G/(G-1)}[\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g$ where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$
  - same as jackknife
- Finite sample Wald tests
  - at least use $T(G-1)$ p-values and critical values and not $\mathcal{N}[0,1]$
  - Example $G = 10$
    - ★ $t = 1.96$ has $p = 0.082$ using $T(9)$ versus $p = 0.05$ using $\mathcal{N}[0,1]$
  - ad hoc reasonable correction used by Stata.

6.3 Solution 2: Cluster Bootstrap with Asymptotic Refinement

- Cameron, Gelbach and Miller (2008)
  - Test $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$ using $w = (\widehat{\beta}_1 - \beta_1^0)/s_{\widehat{\beta}_1}$
  - perform a cluster bootstrap with asymptotic refinement
  - then true test size is $\alpha + O(G^{-3/2})$ rather than usual $\alpha + O(G^{-1})$
  - hopefully improvement when G is small
  - wild cluster percentile-t bootstrap is best
  - better than pairs cluster percentile-t bootstrap .

Wild Cluster Bootstrap

1. Obtain the OLS estimator $\widehat{\boldsymbol{\beta}}$ and OLS residuals $\widehat{\mathbf{u}}_g$, $g = 1, ..., G$.
   - Best to use residuals that impose H_0.
2. Do B iterations of this step. On the b^{th} iteration:
   1. For each cluster $g = 1, ..., G$, form
      $\widehat{\mathbf{u}}^*_g = \widehat{\mathbf{u}}_g$ or $\widehat{\mathbf{u}}^*_g = -\widehat{\mathbf{u}}_g$ each with probability 0.5
      and 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 $s_{\widehat{\beta}^*_{1,b}}$ and given these 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 $w^*_1, ..., w^*_B$.

Current Research

- Webb (2013) proposes using a six-point distribution for the weights d_g in $\widehat{\mathbf{u}}_g^* = d_g \widehat{\mathbf{u}}_g$.
  - The weights d_g have a $1/6$ chance of each value in $\{-\sqrt{1.5}, -\sqrt{1}, -\sqrt{.5}, \sqrt{.5}, \sqrt{1}, \sqrt{1.5}\}$.
  - Works better with few clusters than two-point
    - ★ Two-point cluster gives only 2^{G-1} different bootstrap resamples.
  - Also with few clusters need to enumerate rather than bootstrap.
- MacKinnon and Webb (2013) find that unbalanced cluster sizes worsens few clusters problem.
  - Wild cluster bootstrap does well.

Use the Bootstrap with Caution

- We assume clustering does not lead to estimator inconsistency
  - focus is just on the standard errors.
- We assume that the bootstrap is valid
  - this is usually the case for smooth problems with asymptotically normal estimators and usual rates of convergence.
  - but there are cases where the bootstrap is invalid.
- When bootstrapping
  - always set the seed (for replicability)
  - use more bootstraps than the Stata default of 50
    - ★ for bootstraps without asymptotic refinement 400 should be plenty.
- When bootstrapping a fixed effects panel data model
  - the additional option `idcluster()` must be used
    - ★ for explanation see Stata manual [R] bootstrap: Bootstrapping statistics from data with a complex structure.

Solution 3: Improved T Critical Values

- Suppose all regressors are invariant within clusters, clusters are balanced and errors are i.i.d. normal
  - then $y_{ig} = \mathbf{x}_g'\boldsymbol{\beta} + \varepsilon_{ig} \Longrightarrow \bar{y}_g = \bar{\mathbf{x}}_g'\boldsymbol{\beta} + \bar{\varepsilon}_g$ with $\bar{\varepsilon}_g$ i.i.d. normal
  - so Wald test based on OLS is exactly $T(G-L)$, where L is the number of group invariant regressors.
- Extend to nonnormal errors and group varying regressors
  - asymptotic theory when G is small and $N_g \rightarrow \infty$.
  - Donald and Lang (2007) propose a two-step FGLS RE estimator yields t-test that is $T(G-L)$ under some assumptions
  - Wooldridge (2006) proposes an alternative minimum distance method.

Current Research (continued)

- Imbens and Kolesar (2012)
  - Data-determined number of degrees of freedom for t and F tests
  - Builds on Satterthwaite (1946) and Bell and McCaffrey (2002).
  - Assumes normal errors and particular model for Ω.
  - 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 $\mathbf{G}'\widehat{\mathbf{\cdot}}\mathbf{G}$.
  - Find works better than 2-point Wild cluster bootstrap but they did not impose H_0.

- Carter, Schnepel and Steigerwald (2013)
  - provide asymptotic theory when clusters are unbalanced
  - propose a measure of the effective number of clusters
  - $G^* = G/(1+\delta)$
    - ★ where $\delta = \frac{1}{G}\sum_{g=1}^{G}\{(\gamma_g - \bar{\gamma})^2/\bar{\gamma}^2\}$
    - ★ $\gamma_g = \mathbf{e}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'\dot{}_g\mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_k$
    - ★ $\mathbf{e}_k$ is a $K \times 1$ vector of zeroes aside from 1 in the k^{th} position if $\widehat{\beta} = \widehat{\beta}_k$
    - ★ $\bar{\gamma} = \frac{1}{G}\sum_{g=1}^{G}\gamma_g$.
- Cluster heterogeneity $(\delta \neq 0)$ can arise for many reasons
  - variation in N_g, variation in $\mathbf{X}_g$ and variation in $\dot{}_g$ across clusters.

- Brewer, Crossley and Joyce (2013)
  - Do FGLS as gives both efficiency gains and works well even with few clusters.

6.5 Special Cases

- Bester, Conley and Hansen (2009)
  - obtain $T(G-1)$ in settings such as panel where mixing conditions apply.
- Ibragimov and Muller (2010) take an alternative approach
  - suppose only within-group variation is relevant
  - then separately estimate $\widehat{\boldsymbol{\beta}}_g s$ and average
  - asymptotic theory when G is small and $N_g \longrightarrow \infty$
- A big limitation is assumption of only within variation
  - for example in state-year panel application with clustering on state it rules out $\mathbf{z}_t$ in $y_{st} = \mathbf{x}_{st}'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma} + \varepsilon_{ig}$ where $\mathbf{z}_t$ are for example time dummies.
- This limitation is relevant in DiD models with few treated groups
  - Conley and Taber (2010) present a novel method for that case.

7. Extensions

- The results for OLS and FGLS and t-tests extend to multiple hypothesis tests and IV, 2SLS. GMM and nonlinear estimators.
- These extensions are incorporated in Stata
  - but Stata generally does not use finite-cluster degrees-of-freedom adjustments in computing test p-values and confidence intervals
    - ★ exception is command regress.

Extensions (continued)

- 7.1 Cluster-Robust F-tests
- 7.2 Instrumental Variables Estimators
  - IV, 2SLS, linear GMM
  - Need modified Hausman test for endogeneity : `estat endogenous`
  - Weak instruments:
    - ★ First-stage F-test should be cluster-robust
    - ★ use add-on `xtivreg2`
    - ★ Finlay and Magnusson (2009) have Stata add-on `rivtest.ado`.
- 7.3 Nonlinear Estimators
  - Population-averaged (`xtreg, pa`) and random effects (e.g. `xtlogit, re`) give quite different βs
  - Rarely can eliminate fixed effects if N_g is small.
- 7.4 Cluster-randomized Experiments

8. Empirical Example: Moulton Setting

- Moulton setting
  - Cross-section sample with clustering on state.
- BDM setting
  - Repeated cross-section data with individual data aggregated to state-year.
- Demonstrate
  - the impact of clustering on standard errors and test size
  - and consider various finite-cluster corrections.

8.1 Cross-section individual-level data

- Table 1: Moulton setting.
  - Cross-section individual-level data March 2012 CPS data with state-level regressor and cluster on state.
  - $N = 65685$ and $G = 51$.
  - Compare various standard errors for OLS and FGLS (RE).
- Table 2: 20% subsample of data in Table 1.
  - Now construct a fake dummy and test $H_0 : \beta = 0$.
  - Do this for $G = 50$, 30, 20, 10 and 6
  - $S = 4000$ for $G \leq 10$ and $S = 1000$ for $G > 10$.
  - $B = 399$ (okay for Monte Carlo but set higher in practice).

Table 1 - Cross-section individual level data				
Impacts of clustering and estimator choices on estimated coefficients and standard errors				
		Estimation Method		
		OLS		FGLS (RE)
Slope coefficient		0.0108		0.0314
Standard Errors				
	Default	0.0042		0.0199
	Heteroscedastic Robust	0.0042		-
	Cluster Robust (cluster on State)	0.0229		0.0214
	Pairs cluster bootstrap	0.0224		0.0216
	Number observations	65685		65685
	Number clusters (states)	51		51
	Cluster size range	519 to 5866		519 to 5866
	Intraclass correlation	0.018		-

Notes: March 2012 CPS data, from IPUMS download. Default standard errors for OLS assume errors are iid; default standard errors for FGLS assume the Random Effects model is correctly specified. The Bootstrap uses 399 replications. A fixed effect model is not possible, since the regressor is invariant within states.

Table 2 - Cross-section individual level data

Monte Carlo rejection rates of true null hypothesis (slope = 0) with different number of clusters and different rejection methods

Nominal 5% rejection rates

	Wald test method	Numbers of Clusters				
		6	10	20	30	50
	Different standard errors and critical values					
1	White Robust, T(N-k) for critical value	0.439	0.457	0.471	0.462	0.498
2	Cluster on state, T(N-k) for critical value	0.215	0.147	0.104	0.083	0.078
3	Cluster on state, T(G-1) for critical value	0.125	0.103	0.082	0.069	0.075
4	Cluster on state, T(G-2) for critical value	0.105	0.099	0.076	0.069	0.075
5	Cluster on state, CR2 bias correction, T(G-1) for critical value	0.082	0.070	0.062	0.060	0.065
6	Cluster on state, CR3 bias correction, T(G-1) for critical value	0.048	0.050	0.050	0.052	0.061
7	Cluster on state, CR2 bias correction, IK degrees of freedom	0.052	0.050	0.047	0.047	0.054
8	Cluster on state, T(CSS effective # clusters)	0.114	0.079	0.057	0.056	0.061
9	Pairs cluster bootstrap for standard error, T(G-1) for critical value	0.082	0.072	0.069	0.067	0.074
	Bootstrap Percentile-T methods					
10	Pairs cluster bootstrap	0.009	0.031	0.046	0.051	0.061
11	Wild cluster bootstrap, Rademacher 2 point distribution, low-p-value	0.097	0.065	0.062	0.051	0.060
12	Wild cluster bootstrap, Rademacher 2 point distribution, mid-p-value	0.068	0.065	0.062	0.051	0.060
13	Wild cluster bootstrap, Rademacher 2 point distribution, high-p-value	0.041	0.064	0.062	0.051	0.060
14	Wild cluster bootstrap, Webb 6 point distribution	0.079	0.067	0.061	0.051	0.061
15	Wild cluster bootstrap, Rademacher 2 pt, do not impose null hypothesis	0.086	0.063	0.050	0.053	0.056
16	IK effective DOF (mean)	3.3	5.6	9.4	12.3	16.9
17	IK effective DOF (5th percentile)	2.7	3.7	4.9	6.3	9.6
18	IK effective DOF (95th percentile)	3.8	7.2	14.5	20.8	29.5
19	CSS effective # clusters (mean)	4.7	6.6	9.9	12.7	17
20	Average number of observations	1554	2618	5210	7803	13055

Notes: March 2012 CPS data, 20% sample from IPUMS download. For 6 and 10 clusters, 4000 Monte Carlo replications. For 20-50 clusters, 1000 Monte Carlo replications. The Bootstraps use 399 replications. "IK effective DOF" from Imbens and Kolesar (2013), and "CSS effective # clusters" from Carter, Schnepel and Steigerwald (2013), see Subsection VI.D. Row 11 uses lowest p-value from interval, when Wild percentile-T bootstrapped p-values are not point identified due to few clusters. Row 12 uses mid-range of interval, and row 13 uses largest p-value of interval.

8.2 BDM Setting with repreated c

- Table 3: BDM setting.
  - Panel level state-year data March 1977-2012 CPS data.
  - Aggregated from individual level data using Hansen (2007) method
    - ★ OLS regress y_{its} on regresors $\mathbf{x}_{its}$ and state-year dummies D_{ts} gives coefficients $\widetilde{y}_{ts}$
    - ★ OLS regress $\widetilde{y}_{ts} = \alpha_s + \delta_t + \beta \times d_{ts} + u_{ts}$
  - $G = 51$, $T = 36$, $N = G \times T = 1836$,
  - Compare various standard errors for FE-OLS and FE-FGLS (AR(1)).
- Table 4: Same data as Table 3.
  - Now construct a fake serially correlated dummy and test $H_0 : \beta = 0$.
  - Do this for $G = 50$, 30, 20, 10 and 6.

Table 3 - State-year panel data with differences-in-differences estimation

Impacts of clustering and estimation choices on estimated coefficients, standard errors, and p-values

		Standard Errors			p-values		
	Model:	1	2	3	1	2	3
	Estimation Method:	OLS-FE	OLS-no FE	FGLS AR(1)	OLS-FE	OLS-no FE	FGLS AR(1)
Slope coefficient		0.0156	0.0040	-0.0042			
Standard Errors							
1	Default standard errors, T(N-k) for critical value	0.0037	0.0062	0.0062	0.000	0.521	0.494
2	White Robust, T(N-k) for critical value	0.0037	0.0055	na	0.000	0.470	na
3	Cluster on state, T(G-1) for critical value	0.0119	0.0226	0.0084	0.195	0.861	0.617
4	Cluster on state, CR2 bias correction, T(G-1) for critical value	0.0118	0.0226	na	0.195	0.861	na
5	Cluster on state, CR2 bias correction, IK degrees of freedom	0.0118	0.0226	na	0.195	0.861	na
6	Pairs cluster bootstrap for standard error, T(G-1) for critical value	0.0118	0.0221	0.0086	0.191	0.857	0.624
	Bootstrap Percentile-T methods						
7	Pairs cluster bootstrap	na	na		0.162	0.878	
8	Wild cluster bootstrap, Rademacher 2 point distribution	na	na		0.742	0.968	
9	Wild cluster bootstrap, Webb 6 point distribution	na	na		0.722	0.942	
10	Imbens-Kolesar effective DOF	50	50				
11	C-S-S effective # clusters	51	51				
	Number observations	1836	1836	1836			
	Number clusters (states)	51	51	51			

Notes: March 1997-2012 CPS data, from IPUMS download. Models 1 and 3 include state and year fixed effects, and a "fake policy" dummy variable that turns on in 1995 for a random subset of half of the states. Model 2 includes year fixed effects but not state fixed effects. The Bootstraps use 999 replications. Model 3 uses FGLS, assuming an AR(1) error within each state. "IK effective DOF" from Imbens and Kolesar (2013), and "CSS effective # clusters" from Carter, Schnepel and Steigerwald (2013), see Subsection VI.D.

Table 4 - State-year panel data with differences-in-differences estimation
Monte Carlo rejection rates of true null hypothesis (slope = 0) with different # clusters and different rejection meth
Nominal 5% rejection rates

	Estimation Method	Numbers of Clusters			
		6	10	20	30
	Wald Tests				
1	Default standard errors, T(N-k) for critical value	0.589	0.570	0.545	0.526
2	Cluster on state, T(N-k) for critical value	0.149	0.098	0.065	0.044
3	Cluster on state, T(G-1) for critical value	0.075	0.066	0.052	0.039
4	Cluster on state, T(G-2) for critical value	0.059	0.063	0.052	0.038
5	Pairs cluster bootstrap for standard error, T(G-1) for critical value	0.056	0.060	0.050	0.036
	Bootstrap Percentile-T methods				
6	Pairs cluster bootstrap	0.005	0.019	0.051	0.044
7	Wild cluster bootstrap, Rademacher 2 point distribution	0.050	0.059	0.050	0.036
8	Wild cluster bootstrap, Webb 6 point distribution	0.056	0.059	0.048	0.037

Notes: March 1997-2012 CPS data, from IPUMS download. Models include state and year fixed effects, and a "fake poli
dummy variable that turns on in 1995 for a random subset of half of the states. For 6 and 10 clusters, 4000 Monte Carlo
replications. For 20-50 clusters, 1000 Monte Carlo replications. The Bootstraps use 399 replications.

9. Current research

- Andreas Hagemann (2016), "Cluster-Robust Bootstrap Inference in Quantile Regression Models," JASA, forthcoming.
  - wild cluster bootstrap for quantile regression.
- James G. MacKinnon and Matthew D. Webb (2016), "Randomization Inference for Difference-in-Differences with Few Treated Clusters" http://www.carleton.ca/economics/wp-content/uploads/cep16-11.pdf
  - Randomization and bootstrap methods for differences-in-differences with few clusters.
- James E. Pustejovsky and Elizabeth Tipton (2016), "Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models." http://arxiv.org/pdf/1601.01981v1
  - Imbens and Kolesar extended to multiple hypothesis tests.

Current research (contined)

- Rustam Ibragimov and Ulrich K. Müller (2016), "Inference with Few heterogeneous Clusters," R.E.Stat., 83-96.
  - Extends Ibragimov and Müller (2010) from one-sample t-test to two-sample t-test.
- Alberto Abadie, Susan Athey, Guido W. Imbens, Jeffrey M. Wooldridge (2014), "Finite Population Causal Standard Errors," NBER Working Paper 20325.
  - proposes randomization-based standard errors that in general are smaller than the conventional robust standard errors.
- A. Colin Cameron and Douglas L. Miller (2014), "Robust Inference for Dyadic Data". http://cameron.econ.ucdavis.edu/research/dyadic_cameron_miller_dec
  - robust inference for paired data such as cross-country trade.

10. Conclusion

- Where clustering is present it is important to control for it.
- We focus on obtaining cluster-robust standard errors
  - though clustering may also lead to estimator inconsistency.
- Many Stata commands provide cluster-robust standard errors using option `vce()`
  - a cluster bootstrap can be used when option `vce()` does not include clustering.
- In practice
  - it can be difficult to know at what level to cluster
  - the number of clusters may be few and asymptotic theory is in the number of clusters.