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
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 $\hat{\beta}$ is consistent for $\beta$.

Concerned with getting the correct standard errors of $\hat{\beta}$

- 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

  - CPS individual data on male wages.
  - But there is no individual data on job injury rate.
  - Instead aggregated data on occupation injury rates

- OLS estimate model for individual $i$ in occupation $g$

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

- 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnw</td>
<td>1498</td>
<td>2.455199</td>
<td>.559654</td>
<td>1.139434</td>
<td>4.382027</td>
</tr>
<tr>
<td>occrate</td>
<td>1498</td>
<td>3.208274</td>
<td>2.990179</td>
<td>.461773</td>
<td>10.78546</td>
</tr>
<tr>
<td>potexp</td>
<td>1498</td>
<td>19.91288</td>
<td>11.22332</td>
<td>0</td>
<td>53.5</td>
</tr>
<tr>
<td>potexpsq</td>
<td>1498</td>
<td>522.4017</td>
<td>516.9058</td>
<td>0</td>
<td>2862.25</td>
</tr>
<tr>
<td>educ</td>
<td>1498</td>
<td>12.97296</td>
<td>2.352056</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>union</td>
<td>1498</td>
<td>.1321762</td>
<td>.3387954</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>nonwhite</td>
<td>1498</td>
<td>.1008011</td>
<td>.3011657</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>northe</td>
<td>1498</td>
<td>.2503338</td>
<td>.4333499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>midw</td>
<td>1498</td>
<td>.2683578</td>
<td>.4432528</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>west</td>
<td>1498</td>
<td>.2503338</td>
<td>.4333499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>occ_id</td>
<td>1498</td>
<td>182.506</td>
<td>99.74337</td>
<td>63</td>
<td>343</td>
</tr>
</tbody>
</table>
- 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)nestimates 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 $N(0, 1)$

<table>
<thead>
<tr>
<th>Variable</th>
<th>one_iid</th>
<th>one_het</th>
<th>one_clu</th>
<th>one_xtclu</th>
</tr>
</thead>
<tbody>
<tr>
<td>occrate</td>
<td>-0.0448</td>
<td>-0.0448</td>
<td>-0.0448</td>
<td>-0.0448</td>
</tr>
<tr>
<td></td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.0164</td>
<td>0.0163</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0260</td>
<td>0.0060</td>
</tr>
<tr>
<td>potexp</td>
<td>0.0420</td>
<td>0.0420</td>
<td>0.0420</td>
<td>0.0420</td>
</tr>
<tr>
<td></td>
<td>0.0039</td>
<td>0.0037</td>
<td>0.0073</td>
<td>0.0073</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>potexpsq</td>
<td>-0.0006</td>
<td>-0.0006</td>
<td>-0.0006</td>
<td>-0.0006</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>educ</td>
<td>0.0840</td>
<td>0.0840</td>
<td>0.0840</td>
<td>0.0840</td>
</tr>
<tr>
<td></td>
<td>0.0055</td>
<td>0.0065</td>
<td>0.0175</td>
<td>0.0175</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0000</td>
</tr>
<tr>
<td>union</td>
<td>0.2557</td>
<td>0.2557</td>
<td>0.2557</td>
<td>0.2557</td>
</tr>
<tr>
<td></td>
<td>0.0362</td>
<td>0.0336</td>
<td>0.0892</td>
<td>0.0889</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0210</td>
<td>0.0040</td>
</tr>
</tbody>
</table>
And cluster-robust variance matrix is rank deficient

<table>
<thead>
<tr>
<th></th>
<th>nonwhite</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-0.1057</td>
<td>-0.1057</td>
<td>-0.1057</td>
<td>-0.1057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0391</td>
<td>0.0369</td>
<td>0.0502</td>
<td>0.0501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.007</td>
<td>0.004</td>
<td>0.068</td>
<td>0.035</td>
</tr>
<tr>
<td>northe</td>
<td></td>
<td>0.0501</td>
<td>0.0501</td>
<td>0.0501</td>
<td>0.0501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0326</td>
<td>0.0340</td>
<td>0.0225</td>
<td>0.0224</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.125</td>
<td>0.141</td>
<td>0.057</td>
<td>0.025</td>
</tr>
<tr>
<td>midw</td>
<td></td>
<td>-0.0124</td>
<td>-0.0124</td>
<td>-0.0124</td>
<td>-0.0124</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0319</td>
<td>0.0329</td>
<td>0.0300</td>
<td>0.0299</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.698</td>
<td>0.707</td>
<td>0.691</td>
<td>0.679</td>
</tr>
<tr>
<td>west</td>
<td></td>
<td>0.0402</td>
<td>0.0402</td>
<td>0.0402</td>
<td>0.0402</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0339</td>
<td>0.0347</td>
<td>0.0370</td>
<td>0.0369</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.236</td>
<td>0.246</td>
<td>0.309</td>
<td>0.276</td>
</tr>
<tr>
<td>_cons</td>
<td></td>
<td>0.9679</td>
<td>0.9679</td>
<td>0.9679</td>
<td>0.9679</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0876</td>
<td>0.1014</td>
<td>0.2461</td>
<td>0.2453</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>1498</th>
<th>1498</th>
<th>1498</th>
<th>1498</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_clust rank</td>
<td>10.0000</td>
<td>10.0000</td>
<td>9.0000</td>
<td>8.0000</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>95.2130</td>
<td>89.0902</td>
<td>.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $t$
  \[
y_{ts} = \alpha + x_{ts}' \beta + \gamma \times d_{ts} + u_{ts}.
\]

- 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

\[
y_{its} = \alpha + x'_{its}\beta + \gamma \times d_{ts} + u_{its}
\]

  > 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

1. Introduction
2. Cluster-Robust Inference for OLS
3. Cluster-Specific Fixed Effects
4. What to Cluster Over?
5. Multi-way Clustering
6. Few Clusters
7. Extensions (beyond OLS)
8. Empirical Example
9. Conclusion
2. Cluster-Robust Inference for OLS

- Clustered errors: \( y_{ig} = x'_{ig} \beta + u_{ig} \) 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

\[
\tau_j \approx 1 + \rho_{x_j} \rho_u (\bar{N}_g - 1),
\]

- (1) \( \rho_{x_j} \) is the within cluster correlation of \( x_j \)
- (2) \( \rho_u \) 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 \( V[\hat{\beta}] \) is natural extension of White’s (1980) heteroskedastic-robust estimate
  - but requires number of groups \( G \to \infty \).

- 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 $\mu$ 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 $\text{Var}[y] = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$. Then

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

  $$= \frac{1}{N^2} \left[ N\sigma^2 + N(N-1)\rho\sigma^2 \right] = \frac{1}{N} \sigma^2 \{1 + (n-1)\rho\}.$$
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:

$$y_{ig} = x'_{ig} \beta + u_{ig}, \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G,$$

$$y_g = X_g \beta + u_g, \quad g = 1, \ldots, G,$$

$$y = X \beta + u.$$

- OLS estimator

$$\hat{\beta} = \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} x_{ig} x'_{ig} \right)^{-1} \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} x_{ig} y_{ig} \right)$$

$$= \left( \sum_{g=1}^{G} X_g' X_g \right)^{-1} \left( \sum_{g=1}^{G} X_g' y_g \right)$$

$$= (X'X)^{-1} X'y.$$
2. Cluster-Robust Inference  

2.2 Clustered Errors

As usual

\[ \hat{\beta} = \beta + (X'X)^{-1}X'u \]

\[ = \beta + (X'X)^{-1}(\sum_{g=1}^{G} X_g u_g). \]

Assume independence over \( g \) and correlation within \( g \)

\[ \mathbb{E}[u_{ig} u_{jg'} | x_{ig}, x_{jg'}] = 0, \text{ unless } g = g'. \]

Then \( \hat{\beta} \sim \mathcal{N} [\beta, V[\hat{\beta}]] \) with asymptotic variance

\[ \text{Avar}[\hat{\beta}] = (\mathbb{E}[X'X])^{-1}(\sum_{g=1}^{G} \mathbb{E}[X'_g u_g u'_g X_g])(\mathbb{E}[X'X])^{-1} \]

\[ \neq \sigma^2_u (\mathbb{E}[X'X])^{-1}. \]
Consequences - KEY RESULT FOR INSIGHT

- Suppose equicorrelation within cluster $g$

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

- this arises in a random effects model with $u_{ig} = \alpha_g + \varepsilon_{ig}$, where $\alpha_g$ and $\varepsilon_{ig}$ 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

\[
\tau_j \approx 1 + \rho_{xj} \rho_u (\bar{N}_g - 1),
\]

- (1) $\rho_{xj}$ is the within cluster correlation of $x_j$
- (2) $\rho_u$ 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)

Practice: Moulton (1986, 1990) showed that the variance inflation can be large even if $\rho_u$ is small

- especially with a grouped regressor (same for all individuals in group) so that $\rho_x = 1$.
- CPS data example:
  \[
  N_g = 81, \; \rho_x = 1 \text{ and } \rho_u = 0.1
  \]
  \[
  \tau_j \sim 1 + \rho_x \rho_u (N_g - 1) = 1 + 1 \times 0.1 \times 80 = 9.
  \]
  * 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

\[
\text{Avar}[\hat{\beta}] = \left(\text{E}[X'X]\right)^{-1} \left(\sum_{i=1}^{N} \text{E}[u_i^2 x_i x_i']\right) \left(\text{E}[X'X]\right)^{-1}
\]

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

\[
\hat{V}[\hat{\beta}] = \left(X'X\right)^{-1} \left(\sum_{i=1}^{N} \hat{u}_i^2 x_i x_i'\right) \left(X'X\right)^{-1}.
\]

- Need \( \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 x_i x_i' - \frac{1}{N} \sum_{i=1}^{N} \text{E}[u_i^2 x_i x_i'] \xrightarrow{p} 0 \)
  
  - not \( \hat{u}_i^2 \xrightarrow{p} \text{E}[u_i^2] \)
Similarly for OLS with independent clustered errors

\[ \text{Avar}[\hat{\beta}] = (\mathbb{E}[X'X])^{-1}(\sum_{g=1}^{G} \mathbb{E}[X'_g u_g u'_g X_g])(\mathbb{E}[X'X])^{-1} \]

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

\[ \hat{V}_{CR}[\hat{\beta}] = (X'X)^{-1}(\sum_{g=1}^{G} X'_g \tilde{u}_g \tilde{u}'_g X_g)(X'X)^{-1}. \]

- Stata uses \( \tilde{u}_g = c\hat{u}_g = c(y_g - X_g \hat{\beta}) \) where \( c = \frac{G}{G-1} \frac{N-1}{N-K} \approx \frac{G}{G-1} \).
2.3 The Cluster-Robust Variance Matrix Estimate

- 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 $N_g \to \infty$.
- 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 $\Omega_g = E[u_g u'_g | X_g]$, e.g. within-cluster equicorrelation.
- Given $\hat{\Omega} \xrightarrow{p} \Omega$, the feasible GLS estimator of $\beta$ is

$$\hat{\beta}_{FGLS} = \left( \sum_{g=1}^{G} X'_g \hat{\Omega}_g^{-1} X_g \right)^{-1} \sum_{g=1}^{G} X'_g \hat{\Omega}_g^{-1} y_g.$$ 

- Default $\hat{V}[\hat{\beta}_{FGLS}] = (X'\hat{\Omega}^{-1} X)^{-1}$ requires correct $\Omega$.
- To guard against misspecified $\Omega_g$ use cluster-robust

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

- where $\hat{u}_g = y_g - X_g \hat{\beta}_{FGLS}$ and $\hat{\Omega} = \text{Diag}[\hat{\Omega}_g]$,
- assumes $u_g$ and $u_h$ are uncorrelated, for $g \neq h$
- and needs $G \rightarrow \infty$. 
FGLS Examples

- **Example 1 - Moulton setting**
  - Random effects model: $y_{ig} = x'_{ig} \beta + \alpha_g + \epsilon_{ig}$
    - `xtreg, re vce(robust)`
  - Richer hierarchical linear model or mixed model
    - Stata 13: `mixed, vce(robust)`

- **Example 2 - BDM setting**
  - AR(1) error $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$ and $\epsilon_{it}$ 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 $G \to \infty$
  - Unless FE’s present and $N_g$ small (see later).
2.5 The CRVE can be rank deficient

- $\hat{V}_{CR}[\hat{\beta}]$ can be rank deficient
  - rank is as most minimum of $K$ and $G - 1$
  - $\hat{\beta} = C'C$, where $C' = [X'_1 \hat{u}_1 \cdots X'_G \hat{u}_G]$ is $K \times G$
  - and $X'_1 \hat{u}_1 + \cdots + X'_G \hat{u}_G = 0$

- 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 \( \{(y_1^*, X_1^*), \ldots, (y_G^*, X_G^*)\} \) by resampling with replacement \( G \) times from the original sample of clusters
  - (2) compute \( \hat{\beta}_b \) (estimate of \( \beta \)) in the \( b^{th} \) bootstrap sample.

- Compute the variance of the \( B \) estimates \( \hat{\beta}_1, \ldots, \hat{\beta}_B \) as

\[
\hat{V}_{CR; boot} [\hat{\beta}] = \frac{1}{B - 1} \sum_{b=1}^{B} (\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})',
\]

where \( \bar{\beta} = B^{-1} \sum_{b=1}^{B} \hat{\beta}_b \) 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 \( y_{ig} = x_{ig}' \beta + \alpha_g + u_{ig} = x_{ig}' \beta + \sum_{h=1}^{G} \alpha_g d_{hig} + u_{ig} \).
- 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 \( N_g \) 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 \( \hat{\alpha}_g \) inconsistent
  - Hansen (2007b) provides bias-corrected FGLS for AR(p) errors
  - Brewer, Crossley and Joyce (2013) implement in DiD setting
- 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.

4. What to Cluster Over?

4.2 Clustering Due to Survey Design

- 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

  - 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

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

- What should we do?
  - Ad hoc robust: OLS and robust cluster on industry for $\hat{\gamma}$ and robust cluster on occupation for $\hat{\delta}$.
  - Non-robust: FGLS two-way random effects: $u_{igh} = \varepsilon_{g} + \varepsilon_{h} + \varepsilon_{igh}; \varepsilon_{g}, \varepsilon_{h}, \varepsilon_{igh}$ i.i.d.
  - Two-way robust: next
5.1 Two-way Cluster-Robust

- Robust variance matrix estimates are of the form
  \[ \hat{\text{Avar}}[\hat{\beta}] = (X'X)^{-1}\hat{B}(X'X)^{-1} \]

- For one-way clustering with clusters \( g = 1, \ldots, G \) we can write
  \[ \hat{B} = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in same cluster } g] \]
  - where \( \hat{u}_i = y_i - x_i' \hat{\beta} \) and
  - the indicator function \( 1[A] \) equals 1 if event \( A \) occurs and 0 otherwise.

- For two-way clustering with clusters \( g = 1, \ldots, G \) and \( h = 1, \ldots, H \)
  \[ \hat{B} = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ share any of the two clusters}] \]
  \[ = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in same cluster } g] \]
  \[ + \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in same cluster } h] \]
  \[ - \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in both cluster } g \text{ and } h]. \]
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
  \[
  \hat{V}_{\text{two-way}}[\hat{\beta}] = \hat{V}_G[\hat{\beta}] + \hat{V}_H[\hat{\beta}] - \hat{V}_{G\cap H}[\hat{\beta}],
  \]


- Extends to multi-way clustering.

Early empirical applications that independently proposed this method include Acemoglu and Pischke (2003).
5.2 Implementation

If $\hat{V}[\hat{\beta}]$ is not positive-definite (small $G$, $H$) then

- Decompose $\hat{V}[\hat{\beta}] = U\Lambda U'$; $U$ contains eigenvectors of $\hat{V}$, and $\Lambda = \text{Diag}[\lambda_1, ..., \lambda_d]$ contains eigenvalues.
- Create $\Lambda^+ = \text{Diag}[\lambda_1^+, ..., \lambda_d^+]$, with $\lambda_j^+ = \max(0, \lambda_j)$, and use $\hat{V}^+[\hat{\beta}] = U\Lambda^+ U'$
- 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 $G \to \infty$ and $H \to \infty$ 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 $\gamma$ and one-way cluster on occupation for $\delta$.

- **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
  \[ y_{igh} = x'_{igh} \beta + \alpha_g + \delta_h + \varepsilon_{ig} \text{ with i.i.d. errors} \]
  \[ \text{xtmixed } y \ x \ || \ _\text{all}: \ R.\text{id1} \ || \ \text{id2}: \ , \ mle. \]
  \[ \text{but cannot then get cluster-robust variance matrix} \]

- Hierarchical linear models or mixed models
  \[ y_{ig} = x'_{ig} \beta_g + u_{ig} \]
  \[ \beta_g = W_g \gamma + v_j \text{ where } u_{ig} \text{ and } v_g \text{ are errors.} \]
  \[ \text{see Rabe-Hesketh and Skrondal (2012)} \]
5.4 Spatial Correlation

- Two-way cluster robust related to time-series and spatial HAC.
- In general, $\hat{B}$ in preceding has the form $\sum_i \sum_j w(i, j) x_i' x_j' \hat{u}_i \hat{u}_j$.
  - 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 $T \to \infty$
  - 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.

6. Inference with Few Clusters

- One-way clustering, and focus on the Wald “t-statistic”
  \[ w = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}} \]

- CRVE assumes \( G \to \infty \). What if \( G \) is small?
- At a minimum use CRVE with rescaled error \( \tilde{u}_g = \sqrt{c} \hat{u}_g \)
  - where \( c = \frac{G}{G-1} \) or \( c = \frac{G}{G-1} \times \frac{N-1}{N-k} \approx \frac{G}{G-1} \)
- 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 $\hat{u}$ systematically biased to zero compared to $u$.
  - e.g. OLS with iid normal errors $E[\hat{u}'\hat{u}] = (N - K)\sigma^2$, 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 $\hat{u}_g = \sqrt{c\hat{u}_g}$, already mentioned.
- CR2VE generalizes HC2 for heteroskedasticity
  
  $\tilde{u}_g^* = [I_{N_g} - H_{gg}]^{-1/2}\hat{u}_g$ where $H_{gg} = X_g(X'X)^{-1}X'_g$
  
  gives unbiased CRVE if errors iid normal

- CR3VE generalizes HC3 for heteroskedasticity
  
  $\tilde{u}_g^+ = \sqrt{G/(G-1)}[I_{N_g} - H_{gg}]^{-1}\hat{u}_g$ where $H_{gg} = X_g(X'X)^{-1}X'_g$
  
  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 $w = (\hat{\beta}_1 - \beta_1^0) / \hat{s}_{\beta_1}$
  - perform a cluster bootstrap with asymptotic refinement
  - then true test size is $\alpha + O(\sqrt{G})$ 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 $\hat{\beta}$ and OLS residuals $\hat{u}_g, g = 1, \ldots, G$.
   - Best to use residuals that impose $H_0$.

2. Do $B$ iterations of this step. On the $b^{th}$ iteration:
   - For each cluster $g = 1, \ldots, G$, form $\hat{u}_g^* = \hat{u}_g$ or $\hat{u}_g^* = -\hat{u}_g$ each with probability 0.5 and hence form $\hat{y}_g^* = X'_g \hat{\beta} + \hat{u}_g^*$.
   - This yields wild cluster bootstrap resample $\{(\hat{y}_1^*, X_1), \ldots, (\hat{y}_G^*, X_G)\}$.

3. Calculate the OLS estimate $\hat{\beta}_{1,b}^*$ and its standard error $s_{\hat{\beta}_{1,b}^*}$ and given these form the Wald test statistic $w_b^* = (\hat{\beta}_{1,b}^* - \hat{\beta}_1) / s_{\hat{\beta}_{1,b}^*}$.

4. Reject $H_0$ at level $\alpha$ if and only if
   $$ w < w_{[\alpha/2]}^* \text{ or } w > w_{[1-\alpha/2]}^*, $$
   where $w_{[q]}^*$ denotes the $q^{th}$ quantile of $w_1^*, \ldots, w_B^*$. 
Current Research

- Webb (2013) proposes using a six-point distribution for the weights $d_g$ in $\hat{u}_g^* = d_g \hat{u}_g$.
  - 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 \texttt{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 \( y_{ig} = x_g' \beta + \epsilon_{ig} \implies \bar{y}_g = \bar{x}_g' \beta + \bar{\epsilon}_g \) with \( \bar{\epsilon}_g \) 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 \( N_g \to \infty \).
  - Donald and Lang (2007) propose a two-step FGLS RE estimator yields t-test that is \( T(G - L) \) under some assumptions
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 $\Omega$.
  - Match first two moments of test statistic with first two moments of $\chi^2$.
  - $\nu^* = (\sum_{j=1}^{G} \lambda_j)^2 / (\sum_{j=1}^{G} \lambda_j^2)$ and $\lambda_j$ are the eigenvalues of the $G \times G$ matrix $G^\top G$.
  - 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
- $G^* = \frac{G}{1 + \delta}$
  - where $\delta = \frac{1}{G} \sum_{g=1}^{G} \{ (\gamma_g - \bar{\gamma})^2 / \bar{\gamma}^2 \}$
  - $\gamma_g = e'_k (X'X)^{-1} X'_g \cdot X_g (X'X)^{-1} e_k$
  - $e_k$ is a $K \times 1$ vector of zeroes aside from 1 in the $k^{th}$ position if $\hat{\beta} = \hat{\beta}_k$
  - $\bar{\gamma} = \frac{1}{G} \sum_{g=1}^{G} \gamma_g$.

Cluster heterogeneity ($\delta \neq 0$) can arise for many reasons
- variation in $N_g$, variation in $X_g$ and variation in $\cdot_g$ 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 \( \hat{\beta}_g \)'s and average
  - asymptotic theory when \( G \) is small and \( N_g \to \infty \)

- A big limitation is assumption of only within variation
  - for example in state-year panel application with clustering on state it rules out \( z_t \) in \( y_{st} = x'_{st} \beta + z'_t \gamma + \epsilon_{ig} \) where \( z_t \) 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.
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 $\beta$s
  - 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

<table>
<thead>
<tr>
<th></th>
<th>Estimation Method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>FGLS (RE)</td>
<td></td>
</tr>
<tr>
<td>Slope coefficient</td>
<td>0.0108</td>
<td>0.0314</td>
<td></td>
</tr>
<tr>
<td>Standard Errors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Default</td>
<td>0.0042</td>
<td>0.0199</td>
<td></td>
</tr>
<tr>
<td>Heteroscedastic Robust</td>
<td>0.0042</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Cluster Robust (cluster on State)</td>
<td>0.0229</td>
<td>0.0214</td>
<td></td>
</tr>
<tr>
<td>Pairs cluster bootstrap</td>
<td>0.0224</td>
<td>0.0216</td>
<td></td>
</tr>
<tr>
<td>Number observations</td>
<td>65685</td>
<td>65685</td>
<td></td>
</tr>
<tr>
<td>Number clusters (states)</td>
<td>51</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>Cluster size range</td>
<td>519 to 5866</td>
<td>519 to 5866</td>
<td></td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>0.018</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Nominal 5% rejection rates</th>
<th>Numbers of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald test method</td>
<td>6</td>
</tr>
<tr>
<td>Different standard errors and critical values</td>
<td></td>
</tr>
<tr>
<td>1 White Robust, T(N-k) for critical value</td>
<td>0.439</td>
</tr>
<tr>
<td>2 Cluster on state, T(N-k) for critical value</td>
<td>0.215</td>
</tr>
<tr>
<td>3 Cluster on state, T(G-1) for critical value</td>
<td>0.125</td>
</tr>
<tr>
<td>4 Cluster on state, T(G-2) for critical value</td>
<td>0.105</td>
</tr>
<tr>
<td>5 Cluster on state, CR2 bias correction, T(G-1) for critical value</td>
<td>0.082</td>
</tr>
<tr>
<td>6 Cluster on state, CR3 bias correction, T(G-1) for critical value</td>
<td>0.048</td>
</tr>
<tr>
<td>7 Cluster on state, CR2 bias correction, IK degrees of freedom</td>
<td>0.052</td>
</tr>
<tr>
<td>8 Cluster on state, T(CSS effective # clusters)</td>
<td>0.114</td>
</tr>
<tr>
<td>9 Pairs cluster bootstrap for standard error, T(G-1) for critical value</td>
<td>0.082</td>
</tr>
<tr>
<td>Bootstrap Percentile-T methods</td>
<td></td>
</tr>
<tr>
<td>10 Pairs cluster bootstrap</td>
<td>0.009</td>
</tr>
<tr>
<td>11 Wild cluster bootstrap, Rademacher 2 point distribution, low-p-value</td>
<td>0.097</td>
</tr>
<tr>
<td>12 Wild cluster bootstrap, Rademacher 2 point distribution, mid-p-value</td>
<td>0.068</td>
</tr>
<tr>
<td>13 Wild cluster bootstrap, Rademacher 2 point distribution, high-p-value</td>
<td>0.041</td>
</tr>
<tr>
<td>14 Wild cluster bootstrap, Webb 6 point distribution</td>
<td>0.079</td>
</tr>
<tr>
<td>15 Wild cluster bootstrap, Rademacher 2 pt, do not impose null hypothesis</td>
<td>0.086</td>
</tr>
<tr>
<td>16 IK effective DOF (mean)</td>
<td>3.3</td>
</tr>
<tr>
<td>17 IK effective DOF (5th percentile)</td>
<td>2.7</td>
</tr>
<tr>
<td>18 IK effective DOF (95th percentile)</td>
<td>3.8</td>
</tr>
<tr>
<td>19 CSS effective # clusters (mean)</td>
<td>4.7</td>
</tr>
<tr>
<td>20 Average number of observations</td>
<td>1554</td>
</tr>
</tbody>
</table>

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 repeated 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 regressors $x_{its}$ and state-year dummies $D_{ts}$ gives coefficients $\tilde{y}_{ts}$
  - OLS regress $\tilde{y}_{ts} = \alpha_s + \delta_t + \beta \times d_{ts} + u_{ts}$
- $G = 51, \ T = 36, \ N = G \times T = 1836$,
- 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

<table>
<thead>
<tr>
<th>Model:</th>
<th>Standard Errors</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Estimation Method:</td>
<td>OLS-FE</td>
<td>OLS-no FE</td>
</tr>
<tr>
<td>Slope coefficient</td>
<td>0.0156</td>
<td>0.0040</td>
</tr>
<tr>
<td>Standard Errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Default standard errors, T(N-k) for critical value</td>
<td>0.0037</td>
<td>0.0062</td>
</tr>
<tr>
<td>2 White Robust, T(N-k) for critical value</td>
<td>0.0037</td>
<td>0.0055</td>
</tr>
<tr>
<td>3 Cluster on state, T(G-1) for critical value</td>
<td>0.0119</td>
<td>0.0226</td>
</tr>
<tr>
<td>4 Cluster on state, CR2 bias correction, T(G-1) for critical value</td>
<td>0.0118</td>
<td>0.0226</td>
</tr>
<tr>
<td>5 Cluster on state, CR2 bias correction, IK degrees of freedom</td>
<td>0.0118</td>
<td>0.0226</td>
</tr>
<tr>
<td>6 Pairs cluster bootstrap for standard error, T(G-1) for critical value</td>
<td>0.0118</td>
<td>0.0221</td>
</tr>
<tr>
<td>Bootstrap Percentile-T methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Pairs cluster bootstrap</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>8 Wild cluster bootstrap, Rademacher 2 point distribution</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>9 Wild cluster bootstrap, Webb 6 point distribution</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>10 Imbens-Kolesar effective DOF</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>11 C-S-S effective # clusters</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>Number observations</td>
<td>1836</td>
<td>1836</td>
</tr>
<tr>
<td>Number clusters (states)</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

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 methods

Nominal 5% rejection rates

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Numbers of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td><strong>Wald Tests</strong></td>
<td></td>
</tr>
<tr>
<td>1 Default standard errors, $T(N-k)$ for critical value</td>
<td>0.589</td>
</tr>
<tr>
<td>2 Cluster on state, $T(N-k)$ for critical value</td>
<td>0.149</td>
</tr>
<tr>
<td>3 Cluster on state, $T(G-1)$ for critical value</td>
<td>0.075</td>
</tr>
<tr>
<td>4 Cluster on state, $T(G-2)$ for critical value</td>
<td>0.059</td>
</tr>
<tr>
<td>5 Pairs cluster bootstrap for standard error, $T(G-1)$ for critical value</td>
<td>0.056</td>
</tr>
</tbody>
</table>

| 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 policy" 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

  - wild cluster bootstrap for quantile regression.

  - Randomization and bootstrap methods for differences-in-differences with few clusters.

  - Imbens and Kolesar extended to multiple hypothesis tests.
Current research (continued)

  - Extends Ibragimov and Müller (2010) from one-sample t-test to two-sample t-test.

  - 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".
  - 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.