Robust Inference with Multi-way Clustering

Colin Cameron, Jonah Gelbach, Doug Miller
U.C. - Davis, U. Arizona, U.C. - Davis

February 2010
1.1 Introduction

- Moulton (1986, 1990) and Bertrand, Duflo, and Mullainathan (2004) showed the importance of controlling for clustering.
- Failure to control underestimates OLS standard errors and overstates t statistics.
- Initially use one-way random effects model.
This paper extends to cluster-robust in two (nonnested) dimensions

- Example 1: More than one grouped regressor.
- Example 2: Cross-section unit and time for panel data.

Outline of presentation

- Lengthy discussion of one-way cluster-robust.
- Present method for two-way cluster-robust.
- Simulation and application.
- Conclusion.
1.2 OLS with Cluster Errors

- Model for $G$ clusters with $N_g$ individuals per cluster:

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

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

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

- OLS estimator

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

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

$$= (X'X)^{-1} X'y.$$
OLS with Clustered Errors (continued)

- As usual

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

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

- Assume independence over \( g \) with

\[ u_g \sim [0, \Sigma_g = E[u_g u'_g]]. \]

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

\[ V[\hat{\beta}] = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g \Sigma_g X'_g\right)(X'X)^{-1}. \]
If ignore clustering the default OLS variance estimate should be inflated by approximately

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

where

- \( \rho_{x_j} \) is the within cluster correlation of \( x_j \)
- \( \rho_u \) is the within cluster error correlation
- \( \bar{N}_g \) is the average cluster size.


Moulton (1986, 1990) showed that could be large even if \( \rho_u \) small. e.g. \( N_G = 81, \rho_x = 1 \) and \( \rho_u = 0.1 \) then \( \tau_j = 9! \).

So should correct for clustering - but \( \Sigma_g \) is unknown.
1.3 Random effects approach

- The original way to obtain consistent variance matrix estimate.
- Assume a **random effects (RE) model**

\[ u_{ig} = \alpha_g + \varepsilon_{ig} \]
\[ \alpha_g \sim iid[0, \sigma^2_{\alpha}] \]
\[ \varepsilon_{ig} \sim iid[0, \sigma^2_{\varepsilon}] \]

Then \( \Sigma_g = \sigma^2_u I_{Ng} + \sigma^2_{\alpha} e_{Ng} e'_{Ng} \) and we use

\[ \hat{V}_{RE}[\hat{\beta}] = (X'X)^{-1} \left( \sum_{g=1}^{G} X_g \hat{\Sigma}_g X'_g \right) (X'X)^{-1}, \]

where \( \hat{\Sigma}_g = \hat{\sigma}^2_u I_{Ng} + \hat{\sigma}^2_{\alpha} e_{Ng} e'_{Ng} \), and \( \hat{\sigma}^2_{\varepsilon} \) and \( \hat{\sigma}^2_{\alpha} \) are consistent.

- Weakness is strong distributional assumptions.
1.4 Cluster-Robust Variance Estimates

- The current method to obtain variance estimates.
- The **cluster-robust variance estimate (CRVE)**
  \[
  \hat{V}_{CR}[\hat{\beta}] = (X'X)^{-1}(\sum_{g=1}^{G} X_g \tilde{u}_g \tilde{u}'_g X'_g)(X'X)^{-1},
  \]
  provides a consistent estimator of \( V_{CR}[\hat{\beta}] \) if
  \[
  \text{plim} \frac{1}{G} \sum_{g=1}^{G} X_g \tilde{u}_g \tilde{u}'_g X'_g = \text{plim} \frac{1}{G} \sum_{g=1}^{G} X_g \Sigma_g X'_g.
  \]
- \( \hat{V}_{CR}[\hat{\beta}] \) yields **cluster-robust standard errors**.
- If OLS residuals are used then \( \tilde{u}_g = \tilde{u}_g = y_g - X_g \hat{\beta} \).
  Then \( \hat{V}_{CR}[\hat{\beta}] \) is downwards-biased for \( V[\hat{\beta}] \).
The CRVE was proposed by Liang and Zeger (1986) for grouped data, proposed by Arellano (1987) for FE estimator for short panels (where the grouping is on the individual), popularized by incorporation in Stata as the cluster option (Rogers (1993)).

Bertrand, Duflo and Mullainathan (2004) in a much-cited paper noted that with state-year data many studies were erroneously clustering on state-year (implying independence over time) rather than clustering on just state.
1.5 CRVE with few clusters

- CRVE assumes $G \to \infty$. What if $G$ is small?
- Stata uses $\tilde{u}_g = \sqrt{c} \hat{u}_g$ where $c = \frac{G}{G-1} \times \frac{N-1}{N-k} \sim \frac{G}{G-1}$ and $T(G - 1)$ critical values. Reasonable starting point.
- Donald and Lang (2007) do asymptotic theory when $G$ is small and $N_g \to \infty$. Then a two-step FGLS RE estimator yields t-test that is $T(G - 2)$ under some assumptions.
- Ibragimov and Muller (2010) do asymptotic theory when $G$ is small and $N_g \to \infty$ and only within-group variation is relevant. Then separately estimate $\hat{\beta}_g$'s and average.
- Cameron, Gelbach and Miller (2007) propose cluster bootstraps with asymptotic refinement.
1.6 Wild Cluster Bootstrap-T Procedure

- Test $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$ using $w = (\hat{\beta}_1 - \beta_1^0)/s_{\hat{\beta}_1}$.

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:

   1. 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 $\{(\hat{y}_1^*, X_1), \ldots, (\hat{y}_G^*, X_G^*)\}$.

   2. 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 $w_{b}^* = (\hat{\beta}_{1,b}^* - \hat{\beta}_1)/s_{\hat{\beta}_{1,b}^*}$.

3. 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 $w_1^*, \ldots, w_B^*$. 
1.7 Other work on clustering

- Bhattacharya (2005) combines stratification and clustering, and in a GMM setting.
- Christian Hansen (2007a) shows that even for long panels it is okay to use CRVE with FE estimator (as shown in simulations by Krezde (2002)).
- Christian Hansen (2007b) shows that for short panels with AR(p) errors bias-corrected FGLS does well.
- Cameron and Miller (2010) show that degrees-of-freedom corrections can lead to cluster-robust variance estimates differing substantially in LSDV versus mean-differenced model if T is small.
2.0 Motivating example

  - 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 occupations injury rates for 387 occupations.

- Model estimated is

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

- What should we do?
  - GLS two-way random effects: $u_{igh} = \varepsilon_g + \varepsilon_h + \varepsilon_{igh}$; $\varepsilon_g$, $\varepsilon_h$, $\varepsilon_{igh}$ i.i.d.
  - Strong assumptions.
  - OLS then cluster on industry for $\hat{\gamma}$ and cluster on occupation for $\hat{\delta}$.
  - Ad hoc.
Return to one-way clustering before extending to two-way.

Clustering on group $g$ (for $g = 1, \ldots, G$) with $N_g$ individuals in group $g$:

$$y_{ig} = x_{ig}' \beta + u_{ig}$$

$$E[u_{ig} u_{ig}' | x_{ig}, x_{ig}'] = 0, \text{ for } i \neq j \text{ unless } g = g'.$$

Then informally the estimated asymptotic variance matrix of $\hat{\beta}_{OLS}$, denoted $\text{Avar}[\hat{\beta}]$, is

$$\text{Avar}[\hat{\beta}] = (X'X)^{-1} \hat{\Sigma} (X'X)^{-1}$$

where
\[
\hat{B} = \sum_{g=1}^{G} X'_g \hat{u}_g \hat{u}'_g X_g
\]

\[
= X' \begin{bmatrix}
\hat{u}_1 \hat{u}'_1 & 0 & \cdots & 0 \\
0 & \hat{u}_2 \hat{u}'_2 & \cdots \\
\vdots & \vdots & \ddots & 0 \\
0 & \cdots & \cdots & \hat{u}_G \hat{u}'_G
\end{bmatrix} X
\]

\[
= X' \begin{bmatrix}
\hat{u} \hat{u}' \ast & 0 & \cdots & 0 \\
0 & E_2 & \cdots \\
\vdots & \vdots & \ddots & \\
0 & \cdots & \cdots & E_G
\end{bmatrix} X
\]

\[
= X' (\hat{u} \hat{u}' \ast S^G) X,
\]
With \( g \) denoting first group and \( h \) denoting second group

\[
y_{igh} = x'_{igh} \beta + u_{igh},
\]

\[
E[u_{igh} u_{ig'h'} | x_{igh}, x_{ig'h'}] = 0, \text{ for } i \neq j \text{ unless } g = g' \text{ or } h = h'.
\]

Then use

\[
\hat{B} = X' (\hat{uu}' \ast S^{GH}) X,
\]

where \( S^{GH} \) is an \( N \times N \) indicator matrix with \( ij^{th} \) entry equal to one if the \( i^{th} \) and \( j^{th} \) observation share any cluster and equal to zero otherwise.
Two-way Clustering

- Define three $N \times N$ indicator matrices:
  - $S^G$ with $i$th entry equal to one if the $i^{th}$ and $j^{th}$ observation belong to the same cluster $g \in \{1, 2, \ldots, G\}$
  - $S^H$ with $i$th entry equal to one if the $i^{th}$ and $j^{th}$ observation belong to the same cluster $h \in \{1, 2, \ldots, H\}$
  - $S^{G \cap H}$ with $i$th entry equal to one if the $i^{th}$ and $j^{th}$ observation belong to both the same cluster $g \in \{1, 2, \ldots, G\}$ and the same cluster $h \in \{1, 2, \ldots, H\}$.

- Then
  \[
  S^{GH} = S^G + S^H - S^{G \cap H},
  \]
  so
  \[
  \hat{B} = X' (\hat{u} \hat{u}' \ast S^G) X + X' (\hat{u} \hat{u}' \ast S^H) X - X' (\hat{u} \hat{u}' \ast S^{G \cap H} \ast X),
  \]
Hence

\[
\text{Avar}[\hat{\beta}] = (X'X)^{-1}X'(\hat{\mu}\hat{\mu}' \ast S^G)X(X'X)^{-1} \\
+ (X'X)^{-1}X'(\hat{\mu}\hat{\mu}' \ast S^H)X(X'X)^{-1} \\
- (X'X)^{-1}X'(\hat{\mu}\hat{\mu}' \ast S^{G \cap H})X(X'X)^{-1}.
\]

The three components can be separately computed by
(1) OLS regression of \( y \) on \( X \) with standard errors computed using clustering on \( g \in \{1, 2, \ldots, G\} \);
(2) OLS regression of \( y \) on \( X \) with standard errors computed using clustering on \( h \in \{1, 2, \ldots, H\} \); and
(3) OLS regression of \( y \) on \( X \) with standard errors computed using clustering on \( (g, h) \in \{(1, 1), \ldots, (G, H)\} \)

Stata add-on cgmreg.ado does this.
Two-way Clustering: other research

- This method has been independently proposed by others (we began in 2004).
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 $\Lambda = \text{Diag}[\lambda_1, \ldots, \lambda_d]$ contains eigenvalues.
- Create $\Lambda^+ = \text{Diag}[\lambda_1^+, \ldots, \lambda_d^+]$, with $\lambda_j^+ = \max(0, \lambda_j)$, and use $\hat{V}^+[\hat{\beta}] = U\Lambda^+ U'$.

- Fixed effects in one or both dimensions
  - We do not formally address 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.
2.4 Multi-way clustering

- Extends to multi-way.
- e.g. 7 variance estimates for three-way clustering and $\hat{B}$ may be written as

$$
\left( \hat{B}_{(1,0,0)} + \hat{B}_{(0,1,0)} + \hat{B}_{(0,0,1)} \right) - \left( \hat{B}_{(1,1,0)} + \hat{B}_{(1,0,1)} + \hat{B}_{(0,1,1)} \right) + \hat{B}_{(1,1,1)}
$$

- Extends to m-estimators and GMM
  e.g. For probit three components can be separately computed by
  (1) probit regression with standard errors computed using clustering on $g \in \{1, 2, ..., G\}$;
  (2) probit regression with standard errors computed using clustering on $h \in \{1, 2, ..., H\}$; and
  (3) probit regression with standard errors computed using clustering on $(g, h) \in \{(1, 1), ..., (G, H)\}$
  Then add (1) and (2) and subtract (3).
M-estimator solves
\[ \sum_{i=1}^{N} h_i(\hat{\theta}) = 0. \]

Under standard assumptions, \( \hat{\theta} \) is asymptotically normal with
\[ \hat{V}[\theta] = \hat{A}^{-1}\hat{B}\hat{A}^{-1}, \]
where \( \hat{A} = \sum_i \frac{\partial h_i}{\partial \theta} \bigg|_{\hat{\theta}} \), and \( \hat{B} \) is an estimate of \( V[\sum_i h_i] \).

For independence over \( i \)
\[ \hat{B} = \sum_{i=1}^{N} \hat{h}_i\hat{h}_i'. \]

For one-way clustering
\[ \hat{B} = \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{h}_i\hat{h}_j' I(i, j) \]
where \( I(i, j) = 1 \) if \( i \) and \( j \) are in the same cluster.
For two-way clustering

\[
\hat{B} = \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{h}_i \hat{h}_j l_1(i, j) + \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{h}_i \hat{h}_j l_2(i, j) - \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{h}_i \hat{h}_j l_{12}(i, j)
\]

where \(l_1(i, j) = 1\) if \(i\) and \(j\) are in the same cluster in dimension 1 and \(l_2(i, j) = 1\) if \(i\) and \(j\) are in the same cluster in dimension 2 and \(l_{12}(i, j) = 1\) if \(i\) and \(j\) are in the same cluster in dimensions 1 and 2.
3.1.1 Dgp with no clustering

- Simulations are for one observation per \((h, g)\) combination. So \(HG\) observations and can drop the subscript \(i\).
- Balanced and equal size groups \(H = G = 10, 20, \ldots, 100\) plus some cases \(H \neq G\).
- Dgp is

\[ y_{gh} = \beta_0 + \beta_1 x_{1gh} + \beta_2 x_{2gh} + \alpha_g + \delta_h + \epsilon_{gh} \]

where \(\beta_0 = \beta_1 = \beta_2 = 1\)

and \(x_{ig}, x_{ih}, \alpha_g, \delta_h\) and \(\epsilon_{igh}\) are all iid \(N[0, 1]\).
Dgp with no clustering

- Compute actual rejection rates for 5% two-sided Wald tests of $\beta_1 = 1$ and $\beta_2 = 1$.
- Different standard error estimates lead to different Wald tests.
- Methods used are
  1. Assume iid errors
  2. One-way cluster on $g$
  3. Two-way random effects
  4. Two-way cluster robust
  5. 4. with $T$ critical values.

- All methods should work with no clustering.
Dgp with no clustering

- All methods work fine with $G = H = 100$
- Small sample over-rejection greatest for two-way cluster-robust, though diminished if use $T(\min(G, H) - 1)$ degrees of freedom.
### Table 1
Rejection probabilities for a true null hypothesis

**True model: iid errors**

<table>
<thead>
<tr>
<th>Number of Group 1 Clusters</th>
<th>Number of Group 2 Clusters</th>
<th>Assume iid errors</th>
<th>One-way cluster robust (cluster on group1)</th>
<th>Two-way random effects</th>
<th>Two-way cluster-robust</th>
<th>Two-way cluster-robust, T critical values</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>5.6%</td>
<td>6.7%</td>
<td>9.0%</td>
<td>9.7%</td>
<td>4.4%</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>6.5%</td>
<td>4.8%</td>
<td>7.9%</td>
<td>5.7%</td>
<td>5.9%</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>5.1%</td>
<td>4.8%</td>
<td>5.9%</td>
<td>5.9%</td>
<td>5.1%</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>5.3%</td>
<td>4.8%</td>
<td>5.8%</td>
<td>5.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>5.0%</td>
<td>5.3%</td>
<td>5.7%</td>
<td>6.1%</td>
<td>5.2%</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>4.9%</td>
<td>5.6%</td>
<td>5.7%</td>
<td>5.8%</td>
<td>4.9%</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
<td>4.7%</td>
<td>5.3%</td>
<td>4.3%</td>
<td>5.6%</td>
<td>4.9%</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
<td>4.6%</td>
<td>5.3%</td>
<td>5.2%</td>
<td>6.0%</td>
<td>4.7%</td>
</tr>
<tr>
<td>90</td>
<td>90</td>
<td>5.7%</td>
<td>4.8%</td>
<td>6.2%</td>
<td>5.0%</td>
<td>5.7%</td>
</tr>
<tr>
<td><strong>100</strong></td>
<td><strong>100</strong></td>
<td><strong>4.7%</strong></td>
<td><strong>5.1%</strong></td>
<td><strong>4.9%</strong></td>
<td><strong>5.3%</strong></td>
<td><strong>4.7%</strong></td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>5.6%</td>
<td>4.0%</td>
<td>8.6%</td>
<td>7.3%</td>
<td>5.5%</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>5.2%</td>
<td>5.1%</td>
<td>6.2%</td>
<td>7.3%</td>
<td>5.3%</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>4.8%</td>
<td>5.1%</td>
<td>7.0%</td>
<td>7.7%</td>
<td>4.7%</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>4.1%</td>
<td>3.8%</td>
<td>5.8%</td>
<td>5.0%</td>
<td>3.9%</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>5.2%</td>
<td>5.0%</td>
<td>5.5%</td>
<td>5.4%</td>
<td>5.0%</td>
</tr>
</tbody>
</table>

Note: The null hypothesis should be rejected 5% of the time. Number of monte carlo simulations is 2000, except that for methods 1-3 it is 1000 when G*H > 1600.
3.1.2 Dgp with two-way clustered homoskedastic errors

- Dgp is
  \[ y_{gh} = \beta_0 + \beta_1 x_{1gh} + \beta_2 x_{2gh} + \alpha_g + \delta_h + u_{gh} \]

where now \( u_{gh} = \varepsilon_g + \varepsilon_h + \varepsilon_{gh} \) and each error is iid \( \mathcal{N}[0, 1] \)
and \( x_{1gh} = z_g + z_{gh} \) where each is iid \( \mathcal{N}[0, 1] \)
and \( x_{2gh} \) is similar.

- Then only the two-way methods work.

- Again two-way random effects does best as this is situation it is intended for.
Table 2
Rejection probabilities for a true null hypothesis

<table>
<thead>
<tr>
<th>Number of Group 1 Clusters</th>
<th>Number of Group 2 Clusters</th>
<th>Assume iid errors</th>
<th>One-way cluster robust (cluster on group1)</th>
<th>Two-way random effects</th>
<th>Two-way cluster-robust</th>
<th>Two-way cluster-robust, T critical values</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>23.4%</td>
<td>23.7%</td>
<td>13.7%</td>
<td>33.2%</td>
<td>6.2%</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>34.0%</td>
<td>32.3%</td>
<td>8.6%</td>
<td>42.7%</td>
<td>5.7%</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>39.7%</td>
<td>41.7%</td>
<td>7.4%</td>
<td>50.6%</td>
<td>5.2%</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>47.7%</td>
<td>47.6%</td>
<td>8.7%</td>
<td>55.2%</td>
<td>6.5%</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50.0%</td>
<td>50.4%</td>
<td>6.0%</td>
<td>58.8%</td>
<td>4.9%</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>54.5%</td>
<td>56.3%</td>
<td>6.4%</td>
<td>64.1%</td>
<td>5.6%</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
<td>54.2%</td>
<td>57.6%</td>
<td>5.5%</td>
<td>64.8%</td>
<td>4.8%</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
<td>61.1%</td>
<td>60.9%</td>
<td>6.5%</td>
<td>67.0%</td>
<td>4.9%</td>
</tr>
<tr>
<td>90</td>
<td>90</td>
<td>63.4%</td>
<td>60.7%</td>
<td>5.6%</td>
<td>67.9%</td>
<td>5.0%</td>
</tr>
<tr>
<td><strong>100</strong></td>
<td><strong>100</strong></td>
<td><strong>62.2%</strong></td>
<td><strong>60.4%</strong></td>
<td><strong>5.8%</strong></td>
<td><strong>67.9%</strong></td>
<td><strong>5.3%</strong></td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>49.9%</td>
<td>21.3%</td>
<td>13.3%</td>
<td>33.4%</td>
<td>8.9%</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>50.8%</td>
<td>33.1%</td>
<td>9.8%</td>
<td>44.5%</td>
<td>6.7%</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>63.0%</td>
<td>21.0%</td>
<td>14.1%</td>
<td>31.7%</td>
<td>10.4%</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>62.9%</td>
<td>33.9%</td>
<td>10.0%</td>
<td>43.7%</td>
<td>6.2%</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>63.9%</td>
<td>54.0%</td>
<td>6.6%</td>
<td>60.5%</td>
<td>5.3%</td>
</tr>
</tbody>
</table>

Note: See Table 1.
3.1.3 Dgp with two-way clustered heteroskedastic errors

- Dgp is

\[ y_{gh} = \beta_0 + \beta_1 x_{1gh} + \beta_2 x_{2gh} + \alpha_g + \delta_h + u_{gh} \]

where \( u_{gh} = \varepsilon_g + \varepsilon_h + \varepsilon_{gh} \) and each error is iid \( \mathcal{N}[0, 1] \)
EXCEPT now \( \varepsilon_{gh} \) is \( \mathcal{N}[0, |x_{1gh} \times x_{2gh}|] \)
and \( x_{1gh} = z_g + z_{gh} \) where each is iid \( \mathcal{N}[0, 1] \)
and \( x_{2gh} \) is similar.

- Then only the two-way cluster-robust method works.
### Table 1

Rejection probabilities for a true null hypothesis

**True model:** a random effect common to each group, and a heteroscedastic component.

<table>
<thead>
<tr>
<th>Number of Group 1 Clusters</th>
<th>Number of Group 2 Clusters</th>
<th>Assume independent errors</th>
<th>One-way cluster robust (cluster on group1)</th>
<th>Two-way random effects</th>
<th>Two-way cluster-robust</th>
<th>Two-way cluster-robust, T critical values</th>
<th>Group fixed effects, Two-way cluster-robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>8.0%</td>
<td>7.9%</td>
<td>15.7% 15.9%</td>
<td>18.4% 16.5%</td>
<td>14.5% 12.9%</td>
<td>8.6% 8.9%</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>7.0%</td>
<td>5.4%</td>
<td>9.5% 13.0%</td>
<td>11.9% 10.9%</td>
<td>10.3% 8.8%</td>
<td>7.1% 5.9%</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>5.9%</td>
<td>6.9%</td>
<td>7.0% 9.7%</td>
<td>8.2% 9.2%</td>
<td>7.1% 8.0%</td>
<td>6.2% 6.0%</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>7.6%</td>
<td>8.2%</td>
<td>6.0% 8.8%</td>
<td>7.1% 7.0%</td>
<td>6.3% 6.5%</td>
<td>5.9% 5.7%</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>11.6%</td>
<td>10.5%</td>
<td>6.1% 9.6%</td>
<td>6.4% 6.9%</td>
<td>6.0% 6.4%</td>
<td>4.4% 5.4%</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>8.1%</td>
<td>5.6%</td>
<td>12.9% 12.9%</td>
<td>13.7% 9.8%</td>
<td>9.6% 5.9%</td>
<td>6.1% 6.0%</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>7.6%</td>
<td>7.5%</td>
<td>7.9% 10.5%</td>
<td>9.2% 8.6%</td>
<td>7.6% 6.6%</td>
<td>5.2% 6.7%</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>10.0%</td>
<td>6.4%</td>
<td>10.4% 10.1%</td>
<td>11.3% 10.0%</td>
<td>7.3% 6.8%</td>
<td>7.5% 6.2%</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>11.7%</td>
<td>5.3%</td>
<td>9.2% 10.8%</td>
<td>9.4% 6.4%</td>
<td>7.7% 4.5%</td>
<td>5.1% 6.2%</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>11.2%</td>
<td>8.1%</td>
<td>6.7% 9.9%</td>
<td>6.9% 6.8%</td>
<td>6.1% 6.2%</td>
<td>6.1% 5.2%</td>
</tr>
</tbody>
</table>

**Note:** The null hypothesis should be rejected 5% of the time. Number of monte carlo simulations is 2000.
3.2 Dgp with Time-State Correlation

- Data on 1,358,623 women from 1979-99 CPS.
- Model is
  \[ y_{ist} = \alpha d_{st} + x'_{ist}\beta + \delta_s + \gamma_t + u_{ist}, \]
  where \( d_{st} \) is placebo policy created to be correlated over \( s \) and \( t \):
  - Specifically \( d_{st} = d^s_{st} + 2d^t_{st}; \ s = 1, \ldots, 50 \) ordered states
    where \( d^s_{st} = 0.6d^s_{st-1} + v^s_{st} \), with \( v^s_{st} \) iid \( \mathcal{N}[0, 1] \)
    and \( d^t_{st} = 0.6d^t_{s-1,t} + v^t_{st} \) with \( v^t_{st} \) iid \( \mathcal{N}[0, 1] \), and also independent
    from other variables. Here the index \( s \) ranges
  - Two-way cluster-robust is best when no fixed effects included.
  - No gain when fixed effects included. So here fixed effects are enough.
Table 2
Rejection probabilities for a true null hypothesis
Monte Carlos with micro (CPS) data

<table>
<thead>
<tr>
<th>RHS control variables</th>
<th>quartic in age, 4 education dummies</th>
<th>quartic in age, 4 education dummies, state fixed effects</th>
<th>quartic in age, 4 education dummies, year fixed effects</th>
<th>quartic in age, 4 education dummies, state and year fixed effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error estimator:</td>
<td>91.6%</td>
<td>92.1%</td>
<td>82.2%</td>
<td>79.0%</td>
</tr>
<tr>
<td>Heteroscedasticity robust</td>
<td>19.8%</td>
<td>22.4%</td>
<td>13.1%</td>
<td>13.9%</td>
</tr>
<tr>
<td>One-way cluster robust (cluster on state-by-year cell)</td>
<td>16.2%</td>
<td>17.0%</td>
<td>12.0%</td>
<td>15.0%</td>
</tr>
<tr>
<td>One-way cluster robust (cluster on state)</td>
<td>10.2%</td>
<td>8.9%</td>
<td>8.7%</td>
<td>7.6%</td>
</tr>
<tr>
<td>One-way cluster robust (cluster on year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-way cluster-robust (cluster on state and year)</td>
<td>7.2%</td>
<td>6.9%</td>
<td>7.6%</td>
<td>7.1%</td>
</tr>
</tbody>
</table>

Note: Data come from 1.3 million employed women from the 1979-1999 March CPS. Table reports rejection rates for testing a (true) null hypothesis of zero on the coefficient of fake treatments. The "treatments" are generated as \( t = e_s + 2 e_t \), with \( e_s \) a state-specific autoregressive component and \( e_t \) a year-specific "spatial" autoregressive component. The outcome is also modified by an independent year-specific autoregressive component. See text for details. 2000 Monte Carlo replications
CPS data on male wages $N = 5960$. Separate data on industry and occupation injury rates. 211 injuries and 387 occupations.

Model estimates is

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

Two-way clustering is best.
### Table 3
Replication of Hersch (1998)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Industry Injury Rate</th>
<th>Occupation Injury Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated slope coefficient:</td>
<td>-1.894</td>
<td>-0.465</td>
</tr>
<tr>
<td>Estimated standard errors and p-values:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Default (iid)</td>
<td>(0.415)</td>
<td>{0.0000}</td>
</tr>
<tr>
<td>Heteroscedastic robust</td>
<td>(0.397)</td>
<td>{0.0000}</td>
</tr>
<tr>
<td>One-way cluster on Industry</td>
<td>(0.643)</td>
<td>{0.0032}</td>
</tr>
<tr>
<td>One-way cluster on Occupation</td>
<td>(0.486)</td>
<td>{0.0001}</td>
</tr>
<tr>
<td>Two-way clustering</td>
<td>(0.702)</td>
<td>{0.0070}</td>
</tr>
</tbody>
</table>

Note: Replication of Hersch (1998), pg 604, Table 3, Panel B, Column 4. Standard errors in parentheses. P-values from a test of each coefficient equal to zero in brackets. Data are 5960 observations on working men from the Current Population Survey. Both columns come from the same regression. There are 211 industries and 387 occupations in the data set.
Dyadic data: trade flows between pairs of countries.
First column of Table 3 of Rose and Engel (2002).
  - gravity model is fitted for the natural logarithm of bilateral trade
  - single cross-section on trade flows between 98 countries with 3262 unique country pairs.
Cameron and Golotvina (2005) control for two-way clustering, using FGLS estimation based on iid country random effects.
  - Here instead apply the more robust method to their paper.
Focus on coefficient of the log product of real GDP (\(= 0.867\))
  - heteroskedastic-robust standard error of 0.013
  - average one-way clustered standard error of 0.031
  - two-way robust standard error is 0.043
  - if country specific effects are included as alternative way to control for the clustering, then the coefficient of the log product of real GDP is no longer identified.
1980-90 March CPS data on 39,063 men.
- Model of retirement ($y_{ist}$) and health insurance coverage ($d_{st}$)

$$
\Pr[y_{ist} = 1] = \Phi(\alpha d_{st} + x_{ist}' \beta + \delta_s + \gamma_t).
$$

- Rotating design - many individuals appear in sample twice.
### Table 6
**Replication of Gruber and Madrian (1995)**

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em><em>Estimated slope coefficient (</em> 1000):</em>*</td>
<td>13.264</td>
<td>1.644</td>
</tr>
<tr>
<td>Estimated standard errors (* 1000):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Default (iid)</td>
<td>(5.709)</td>
<td>(0.675)</td>
</tr>
<tr>
<td>Heteroscedastic robust</td>
<td>(5.732)</td>
<td>(0.684)</td>
</tr>
<tr>
<td>One-way cluster on state-year</td>
<td>(6.265)</td>
<td>(0.759)</td>
</tr>
<tr>
<td>One-way cluster on household id</td>
<td>(5.866)</td>
<td>(0.702)</td>
</tr>
<tr>
<td>One-way cluster on hhid-by-state-year</td>
<td>(5.732)</td>
<td>(0.685)</td>
</tr>
<tr>
<td>Two-way clustering</td>
<td>(6.389)</td>
<td>(0.775)</td>
</tr>
<tr>
<td>One-way cluster on State</td>
<td>(6.030)</td>
<td>(0.718)</td>
</tr>
</tbody>
</table>

Conclusion

- Method is straightforward to implement.
- Works well except in smallest designs with $G = H = 10$.
- Many potential applications.
- Sometimes has big impact and sometimes modest.
- But researcher will not know this a priori.
Asymptotic theory

- Key is

\[ B_0 = \lim \ E \left[ G^{-1} H^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} z_i(\theta)z_j(\theta)' \right]. \]

where \( \sum_{i=1}^{N} z_i(\theta) = \sum_{g=1}^{G} \sum_{h=1}^{H} z_{gh}(\theta) \) and \( z_{gh}(\theta) = \sum_{i \in C_{gh}} z_{igh}(\theta) = \sum_{i \in C_{gh}} (y_{igh} - x'_{igh}\beta)x_{igh}. \)

- For two-way clustering since \( \sum_i z_i(\theta) = \sum_g \sum_h z_{gh}(\theta) \) we have

\[
\begin{align*}
E \left[ G^{-1} H^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} z_i(\theta)z_j(\theta)' \right] &= E \left[ G^{-1} H^{-2} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{g'=1}^{G} \sum_{h'=1}^{H} z_{gh}(\theta)z_{g'h'}(\theta)' \right] \\
&= G^{-1} H^{-2} \sum_{g} \sum_{h} \sum_{h'} E[z_{gh}z'_{gh}] \quad g = g' \\
&\quad + G^{-1} H^{-2} \sum_{h} \sum_{g} \sum_{g'} E[z_{gh}z'_{g'h}] \quad h = h' \\
&\quad - G^{-1} H^{-2} \sum_{g} \sum_{h} E[z_{gh}z'_{gh}].
\end{align*}
\]
Assume the triple sums are of order $GH^2 = G$.
The triple sums are instead of order $GH$, rather than $GH^2$, if the dependence of observations in common cluster $g$ goes to zero as clusters $h$ and $h'$ become further apart.

- case with declining time series dependence or spatial dependence.
- then in $B_0$ normalize by $(GH)$
- rate of convergence of the estimator becomes a faster $\sqrt{GH}$ rather than $\sqrt{G}$.
- Still obtain the same asymptotic variance matrix for $\hat{\beta}$. 


Some References


