Robust Inference with Clustered Data

Colin Cameron
Univ. of California - Davis

Mexico Stata Users Group Meeting
Mexico City May 12, 2011


Mexico Stata Users Group
Abstract

This presentation studies robust inference for regression models where data are clustered, with correlation of observations in the same cluster (such as state) and independence across clusters. The talk ranges from the basics through to complications such as a small number of clusters and two-way clustering. The relevant Stata commands and Stata add-ons, where available, are presented.
1. Introduction

- Failure to control for clustering in OLS regression
  - underestimates OLS standard errors and overstates t statistics.

- Moulton (1986, 1990) and Bertrand, Duflo & Mullainathan (2004) showed
  - the practical importance of controlling for clustering
  - clustering can arise in a wider range of settings than obvious.

- To control for clustering
  - originally use a restrictive one-way random effects model
  - now use cluster-robust standard errors
  - Wooldridge (2003, 2006) and Cameron and Miller (2011) provide surveys.
Outline

1. Introduction
2. Clustering and its Consequences for OLS
3. Cluster-Robust Inference for OLS
4. Inference with Few Clusters
5. Multi-way Clustering
6. Feasible GLS
7. Nonlinear and Instrumental Variables Estimators
8. Stata Implementation
9. Conclusion
2. Clustering and its consequences

- 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
  
  \[ \hat{\beta} = \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} x_{ig} x_{ig}' \right)^{-1} \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} x_{ig} y_{ig} \right) \]
  
  \[ = \left( \sum_{g=1}^{G} X_g' X_g \right)^{-1} \left( \sum_{g=1}^{G} X_g y_g \right) \]
  
  \[ = (X'X)^{-1} X'y. \]
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 \( g \) and correlation within \( g \)

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

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

\[
\text{Avar}[\hat{\beta}] = (E[X'X])^{-1}\left(\sum_{g=1}^{G} E[X'_g u_g u'_g X'_g]\right)(E[X'X])^{-1} \\
\neq \sigma^2_u (E[X'X])^{-1}.
\]
Equicorrelated Errors

- 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

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

- $\rho_{xj}$ 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 the 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: $N_g = 81$, $\rho_x = 1$ and $\rho_u = 0.1$ then $\tau_j \approx 1 + \rho_x \rho_u (\hat{N}_g - 1) = 1 + 1 \times 0.1 \times 80 = 9$.  
  
  - true standard errors are three times the default!

So should correct for clustering even in settings where not obviously a problem.
Panel data

- A second way that clustering can occur is panel data
  - independence across individuals is assumed but correlation over time for a given individual
  - note that here the cluster group $g$ is the individual $i$.

- Equicorrelation is a less reasonable assumption
  - instead correlation decreases with time separation
  - the OLS variance inflation factor may be less than the above
  - but it is still substantial.
3. Cluster-Robust Inference for OLS

- Recall for OLS with independent heteroskedastic errors

\[ \text{Avar}[\hat{\beta}] = (E[XX'])^{-1} (\sum_{i=1}^{N} E[u_i^2 x_i x'_i]) (E[XX'])^{-1} \]

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

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

- Similarly for OLS with independent clustered errors

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

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

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

- Stata uses \( \tilde{u}_g = c \hat{u}_g = c(y_g - X_g \hat{\beta}) \) where \( c = \frac{G}{G-1} \frac{N-1}{N-K} \sim \frac{G}{G-1}. \)
The CRVE was

- proposed by Liang and Zeger (1986) for grouped data
- proposed by Arellano (1987) for the fixed effects estimator for short panels (where the grouping is on the individual)
- popularized by incorporation in Stata as the cluster option (Rogers (1993)).
- also allows for heteroskedasticity so is cluster- and heteroskedastic-robust.
Specifying the clusters

- 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.
3. Cluster-Robust Inference for OLS

Cluster-specific fixed effects

- A cluster-specific fixed effects (FE) model allows a different intercept for each group

\[ y_{ig} = \alpha_i + x_{ig}' \beta + u_{ig} \]

  - in principle, \( \alpha_i \) can account for much of the within-group error correlation
  - in practice, it does not account for all of it.

- The FE estimator is OLS of \((y_{ig} - \bar{y}_i)\) on \((x_{ig} - \bar{x}_i)\)

  - use cluster-robust standard errors after `xtreg, fe`
  - Arellano (1987) showed okay with fixed effects if \( N_g \) fixed and \( G \to \infty \).
  - Kézdi (2004) showed okay for \( N_g \) large relative to \( G \to \infty \).
  - Cameron and Miller (2010) show that if \( N_g \) is small then degrees-of-freedom corrections can lead to cluster-robust variance estimates differing substantially in LSDV versus mean-differenced model.

- Random effects are considered later.
Many observations per cluster

- To date assumed $N_g$ fixed and $G \to \infty$.
- Hansen (2007a) considers case where $N_g \to \infty$ and $G \to \infty$
  - Rate of convergence is $\sqrt{G}$ if there is no within group mixing such as with equicorrelation
  - Rate of convergence is $\sqrt{N_g G}$ if there is within group mixing such as with panel data time series dampening
  - In either case the same asymptotic normal result holds.
Survey design with clustering and stratification

- 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.
- Complex survey data are also stratified
  - this improves estimator efficiency somewhat
  - we ignore this here and focus on clustering.
- And survey data are weighted
  - we ignore assuming conditioning on $x$ handles weighting
- Stata survey commands correct standard errors for both clustering and stratification
4. Inference with few clusters

- CRVE assumes $G \rightarrow \infty$. What if $G$ is small?
  - often still have statistical significance of coefficients
  - but need finite sample corrections to standard errors and tests.

- Finite sample corrected standard errors
  - simplest is $\tilde{u}_g = \sqrt{c} \hat{u}_g$ where $c = \frac{G}{G-1}$ or $c = \frac{G}{G-1} \times \frac{N-1}{N-k} \sim \frac{G}{G-1}$
  - or $\tilde{u}_g^* = [I_{N_g} - H_{gg}]^{-1/2} \hat{u}_g$ where $H_{gg} = X_g (X'X)^{-1} X'_g$
  - or $\tilde{u}_g^+ = \sqrt{G / (G - 1)} [I_{N_g} - H_{gg}]^{-1} \hat{u}_g$
  - these last two are cluster analogs of HC2 and HC3 corrections for heteroskedasticity.

- Finite sample Wald tests
  - at least use $T(G - 1)$ p-values and critical values and not $\mathcal{N}[0, 1]$
  - $G = 10$: $t = 1.96$ has $p = 0.082$ using $T(9)$ versus $p = 0.05$ using $\mathcal{N}[0, 1]$
  - ad hoc reasonable correction used by Stata.
T distribution for inference

- Suppose all regressors are invariant within clusters, clusters are balanced and errors are i.i.d. normal
  - then $y_{ig} = x'_g \beta + \varepsilon_{ig} \Longrightarrow \bar{y}_g = \bar{x}'_g \beta + \bar{\varepsilon}_g$ 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
  - 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 $y_{st} = x'_{st}\beta + z'_t\gamma + \varepsilon_{ig}$ where $z_t$ are for example time dummies.

This limitation is relevant in difference-in-differences models with few treated groups

- Conley and Taber (2010) present a novel method for that case.
Cluster bootstrap with asymptotic refinement

- Cameron, Gelbach and Miller (2007)
  - Test $H_0: \beta_1 = \beta_1^0$ against $H_a: \beta_1 \neq \beta_1^0$ using $w = (\hat{\beta}_1 - \beta_1^0) / s_{\hat{\beta}_1}$
  - 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 bootstrap is best.
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:
   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
      \[
      w^*_b = \frac{(\hat{\beta}^*_1, b - \hat{\beta}_1)}{s_{\hat{\beta}^*_1, b}}.
      \]

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 $w^*_1, \ldots, w^*_B$. 
5. Two-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
Two-way clustering

- Robust variance matrix estimates are of the form

\[
\hat{\text{Var}}[\hat{\beta}] = (X'X)^{-1} \hat{B} (X'X)^{-1}
\]

- 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 \( h = 1, \ldots, H \)

\[
\hat{B} \quad = \quad \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ share any of the two clusters}]
\]

\[
\quad = \quad \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in same cluster } g]
\]

\[
\quad \quad + \quad \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in same cluster } h]
\]

\[
\quad \quad - \quad \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j' \hat{u}_i \hat{u}_j 1[i, j \text{ in both cluster } g \text{ and } h].
\]
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

\[ \hat{\mathbf{V}}_{\text{two-way}}[\hat{\mathbf{\beta}}] = \hat{\mathbf{V}}_G[\hat{\mathbf{\beta}}] + \hat{\mathbf{V}}_H[\hat{\mathbf{\beta}}] - \hat{\mathbf{V}}_{G \cap H}[\hat{\mathbf{\beta}}], \]


- Extends to multi-way clustering.

Early empirical applications that independently proposed this method include Acemoglu and Pischke (2003), and Fafchamps and Gubert (2007).
Practical Considerations

- 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
  - Also Stata add-on xtivreg2.ado has two-way clustering.

- 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 of errors.
CPS data on male wages $N = 5960$.
Separate data on industry and occupation injury rates.
211 injuries and 387 occupations.

Model estimated is

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

Ideally cluster on both industry and occupation.
5. Two-way clustering

- One-way clustering inflates standard errors by 50-60%.
- Two-way clustering inflates by further 10% for industry injury rate.

### 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>(0.235)</td>
</tr>
<tr>
<td>Heteroscedastic robust</td>
<td>(0.397)</td>
<td>(0.260)</td>
</tr>
<tr>
<td>One-way cluster on Industry</td>
<td>(0.643)</td>
<td>(0.251)</td>
</tr>
<tr>
<td>One-way cluster on Occupation</td>
<td>(0.486)</td>
<td>(0.363)</td>
</tr>
<tr>
<td>Two-way clustering</td>
<td>(0.702)</td>
<td>(0.357)</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.
Spatial correlation

- Two-way cluster robust related to time-series and spatial HAC.
- In general, $\hat{\mathbf{B}}$ 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.

- Hybrid estimators combine elements of cluster-robust and HAC:
  - Driscoll and Kraay (1998): each time period is a cluster, plus different time periods may be correlated for a finite time difference with $T \to \infty$.
  - Foote (2007) contrasts various variance matrix estimators in a macroeconomics example.
  - Petersen (2009) contrasts methods for panel data on financial firms.
6. Feasible GLS with Cluster-Robust Inference

- Potential efficiency gains for feasible GLS compared to OLS.
- Specify a model for \( \Omega_g = E[u_g u'_g | X_g] \), such as within-cluster equicorrelation.
- Given a consistent estimate \( \hat{\Omega} \) of \( \Omega \), the feasible GLS estimator of \( \beta \) is

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

- To guard against misspecified \( \Omega_g \) uses cluster-robust variance estimate

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

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 \( G \to \infty \).
Examples of FGLS

- Random effects
  - One-way: \( y_{ig} = x'_{ig} \beta + \alpha_g + \varepsilon_{ig} \) where \( \alpha_g \) and \( \varepsilon_{ig} \) are i.i.d. errors.
  - Two-way: generalizes to \( y_{igh} = x'_{igh} \beta + \alpha_g + \delta_h + \varepsilon_{ig} \) with i.i.d. errors
    - but cannot then get cluster-robust variance matrix.

- Hierarchical linear models or mixed models: \( y_{ig} = x'_{ig} \beta_g + u_{ig} \) and \( \beta_g = W_g \gamma + v_j \) where \( u_{ig} \) and \( v_g \) are errors.

- Time series correlation for panel data
  - Kiefer (1980) \( \Omega_g = \Omega \) and \( \hat{\Omega}_{ij} = G^{-1} \sum_{g=1}^{G} \hat{u}_{ig} \hat{u}_{jg} \), where \( \hat{u}_{ig} \) are OLS residuals
  - AR(p) model for errors
    - Hansen (2007b) bias corrects for FE models

7. Nonlinear estimators: population-averaged models

- Generalized estimating equations (GEE) approach
  - generalizes feasible GLS for clustered data to generalized linear models (GLM)
  - due to Liang and Zeger (1986).

- Method:
  - specify a conditional mean function \( E[y_{ig} | x_{ig}] = m(x_{ig}' \beta) \) e.g. logit
  - specify a variance function and within cluster correlation matrix e.g. equicorrelation to yield the so-called working variance matrix
  - then do analog of feasible GLS
  - get cluster-robust variance matrix that guards against misspecified working variance matrix.
  - asymptotic theory requires that \( G \to \infty \).
Can extend this approach to ML

- specify a conditional density \( f(y_{ig} | x_{ig}) \)
- do quasi-MLE with sandwich variance matrix estimate that controls for clustering
- requires that conditional density \( f(y_{ig} | x_{ig}) \) still correct when clustering is present.
Cluster-specific effects models

- Introduce cluster-specific effects $\alpha_g$ to control for within cluster correlation.
  - Parametric models: density specified for $f(y_{ig} | x_{ig}, \beta, \alpha_g)$, $\alpha_g$ not observed.
  - Conditional mean models: $E[y_{ig} | x_{ig}, \alpha_g] = g(x'_{ig} \beta + \alpha_g)$, $\alpha_g$ not observed.

- Fixed effects approach: $\alpha_g$ are parameters to be estimated
  - Incidental parameters problem if asymptotics $N_g$ is fixed while $G \to \infty$
    - There are $N_g$ parameters $\alpha_1, ..., \alpha_G$ to estimate and $G \to \infty$
    - In general this contaminates estimation of $\beta$ so that $\hat{\beta}$ is a inconsistent.
    - Notable exceptions are logit model, Poisson and nonlinear regression model with additive error.

- Random effects: $\alpha_g$ has density $h(\alpha_g | \eta)$
  - Integrate out $\alpha_g$
  - Often no analytical solution but numerical methods work well as just a one-dimensional integral
  - Results fragile to assumption about distribution of $\alpha_g$ (often $\mathcal{N}[0, \sigma^2_\alpha]$).
Instrumental variables

- Cluster-robust formula is easily adapted to instrumental variables estimation
  - In the just-identified case with \( \widehat{\beta}_{IV} = (Z'X)^{-1}Z'y \) use
    \[
    \widehat{V}_{CR}[\widehat{\beta}_{IV}] = (Z'X)^{-1}(\sum_{g=1}^{G} Z_g \tilde{u}_g \tilde{u}_g' Z_g')(X'X)^{-1}
    \]

  - Shore-Sheppard (1996) showed the importance of doing so, extending Moulton to the IV setting.


- The weak instruments literature focuses on the non-clustered case, often with i.i.d. errors.
  - Finlay and Magnusson (2009) present a test for the significance of the instruments in the reduced form that is cluster robust
  - Stata add-on rivtest.ado.
Cluster-robust extends simply to GMM.

For the $g^{th}$ cluster the moment condition is

- $E[h_g(w_g, \theta)] = 0$ where $w_g$ denotes all variables in the cluster
- assume independence across clusters and $G \to \infty$.

GMM estimator $\hat{\theta}_{GMM}$ minimizes $(\sum_g h_g)' W (\sum_g h_g)$, where $h_g = h_g(w_g, \theta)$.

The variance matrix estimate is

$$\hat{V}[\hat{\theta}_{GMM}] = (\hat{A}' W \hat{A})^{-1} \hat{A}' W \hat{B} W \hat{A} (\hat{A}' W \hat{A})^{-1}$$

- $\hat{A} = \sum_g \partial h_g / \partial \theta'|_{\hat{\theta}}$
- $\hat{B} = \sum_g \hat{h}_g \hat{h}_g'$ for cluster-robust variance matrix estimate.

Bhattacharya (2005) considers stratification in addition to clustering for the GMM estimator.
8. Stata implementation: option vce()

- For cross-section estimation commands
  - use option `vce(cluster cid)` where variable `cid` identifies the clusters.

- For panel estimation commands that begin with `xt`
  - use option `vce(robust)`
  - this gives cluster-robust with clustering on the individual declared in `xtset`
  - same as option `vce(cluster id)` if available
  - in some cases you may want to cluster at a higher level than the individual, such as state.

- For two-way clustering in linear cross-section estimation
  - for OLS use Stata add-on `cgmreg.ado`
  - for IV use Stata add-on `ivreg2.ado`

- For weak instruments test use Stata add-on `rivtest.ado`
Cluster bootstrap

- For commands without a cluster option it may be possible to do a cluster bootstrap.
- A cluster bootstrap resamples with replacement over clusters (rather than over individual observations)
  - for many commands use option \texttt{vce(bootstrap, cluster(cid))}
  - asymptotically equivalent to use option \texttt{vce(cluster cid)}
  - equivalently can use prefix command \texttt{bootstrap, cluster(cid)}:
Bootstrap using prefix command bootstrap, cluster(cid) when option vce(cluster cid) is unavailable.

Example: command xtmixed provides only default standard errors
  ▶ these require correct specification of the functional form for the (clustered) error variance matrix
  ▶ but the estimator is still consistent if this is relaxed
  ▶ so get cluster-robust standard errors using

    bootstrap _b, cluster(cid) reps(400) seed(10101): xtmixed

Cameron and Trivedi (2009, section 13.4) provide other examples
  ▶ bootstrap a user-written program for a two-step estimator
  ▶ bootstrap to implement a robust version of the Hausman test
  ▶ these examples for regular bootstrap can be adapted to cluster bootstrap by adding option cluster(cid).
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.
9. 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 \( \text{vce()} \)
  - a cluster bootstrap can be used when option \( \text{vce()} \) 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.
References


Cameron, A. C., and P. K. Trivedi (2009), Microeconometrics using Stata, College Station, TX: Stata Press.


