Robust Inference with Clustered Data

Colin Cameron
Univ. of California - Davis

Mexico Stata Users Group Meeting
Mexico City May 12, 2011

This talk is based on A. C. Cameron and D. L. Miller (2011), "Robust Inference with Clustered Data", in A. Ullah and D. E. Giles eds., Handbook of Empirical Economics and Finance, CRC Press, pp.1-28.

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
  - White (1984), Liang and Zeger (1986), Arellano (1987), Rogers (1993)
  - Wooldridge (2003, 2006) and Cameron and Miller (2001) 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:

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

- OLS estimator

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

OLS with Clustered Errors (continued)

- As usual

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

- Assume independence over g and correlation within g

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

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

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

Equicorrelated Errors

- Suppose equicorrelation within cluster g

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

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

- The incorrect default OLS variance estimate should be inflated by

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

  - ρ_{x_j} is the within cluster correlation of x_j
  - ρ_u is the within cluster error correlation
  - $\bar{N}_g$ is the average cluster size.
  - Kloek (1981), Scott and Holt (1982).

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

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 rule
  - but it is still substantial.

3. Cluster-Robust Inference for OLS

- Recall for OLS with independent heteroskedastic errors

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

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

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

- Similarly for OLS with independent clustered errors

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

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

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

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

- The CRVE was
  - proposed by 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.

Cluster-specific fixed effects

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

$$y_{ig} = \alpha_i + \mathbf{x}_{ig}'\boldsymbol{\beta} + u_{ig}$$

  - in principle α_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 $(\mathbf{x}_{ig} - \bar{\mathbf{x}}_i)$
  - use cluster-robust standard errors after `xtreg, fe`
  - Arellano (1987) showed okay with fixed effects if N_g fixed and $G \longrightarrow \infty$.
  - Kézdi (2004) showed okay for N_g large relative to $G \longrightarrow \infty$.
  - 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 $N_g \longrightarrow \infty$ and G *fixed*.
- Hansen (2007a) considers case where $N_g \longrightarrow \infty$ and $G \longrightarrow \infty$
  - 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.
- Stata survey commands correct standard errors for both clustering and stratification
  - Bhattacharya (2005) gives a general GMM treatment.

4. Inference with few clusters

- CRVE assumes $G \longrightarrow \infty$. 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 $\widetilde{\mathbf{u}}_g = \sqrt{c}\widehat{\mathbf{u}}_g$ where $c = \frac{G}{G-1}$ or $c = \frac{G}{G-1} \times \frac{N-1}{N-k} \simeq \frac{G}{G-1}$
  - or $\widetilde{\mathbf{u}}_g^* = [\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1/2}\widehat{\mathbf{u}}_g$ where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$
  - or $\widetilde{\mathbf{u}}_g^+ = \sqrt{G/(G-1)}[\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g$
  - 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 $y_{ig} = \mathbf{x}_g'\boldsymbol{\beta} + \varepsilon_{ig} \Longrightarrow \bar{y}_g = \bar{\mathbf{x}}_g'\boldsymbol{\beta} + \bar{\varepsilon}_g$ with $\bar{\varepsilon}_g$ i.i.d. normal
  - so Wald test based on OLS is exactly $T(G-L)$, where L is the number of group invariant regressors.
- Extend to nonnormal errors and group varying regressors
  - asymptotic theory when G is small and $N_g \longrightarrow \infty$.
  - 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 $\widehat{\boldsymbol{\beta}}_g s$ and average
  - asymptotic theory when G is small and $N_g \longrightarrow \infty$
- A big limitation is assumption of only within variation
  - for example in state-year panel application with clustering on state it rules out $\mathbf{z}_t$ in $y_{st} = \mathbf{x}_{st}'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma} + \varepsilon_{ig}$ where $\mathbf{z}_t$ are for example time dummies.
- This limitation is relevant in 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 $w = (\widehat{\beta}_1 - \beta_1^0)/s_{\widehat{\beta}_1}$
  - perform a cluster bootstrap with asymptotic refinement
  - then true test size is $\alpha + O(G^{-3/2})$ rather than usual $\alpha + O(G^{-1})$
  - hopefully improvement when G is small
  - wild cluster bootstrap is best.

Wild cluster bootstrap

1. Obtain the OLS estimator $\widehat{\boldsymbol{\beta}}$ and OLS residuals $\widehat{\mathbf{u}}_g$, $g = 1, ..., G$.
   - Best to use residuals that impose H_0.
2. Do B iterations of this step. On the b^{th} iteration:
   1. For each cluster $g = 1, ..., G$, form
      $\widehat{\mathbf{u}}_g^* = \widehat{\mathbf{u}}_g$ or $\widehat{\mathbf{u}}_g^* = -\widehat{\mathbf{u}}_g$ each with probability 0.5
      and hence form $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g'\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}_g^*$.
      This yields wild cluster bootstrap resample $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$.
   2. Calculate the OLS estimate $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ and given these form the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \widehat{\beta}_1)/s_{\widehat{\beta}_{1,b}^*}$.
3. Reject H_0 at level α if and only if

$$w < w_{[\alpha/2]}^* \text{ or } w > w_{[1-\alpha/2]}^*,$$

where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

5. Two-way clustering

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

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

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

Two-way clustering

- Robust variance matrix estimates are of the form

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

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

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

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

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

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

- Obtain three different cluster-robust "variance" matrices for the estimator by
  - one-way clustering in, respectively, the first dimension, the second dimension, and by the intersection of the first and second dimensions
  - add the first two variance matrices and, to account for double-counting, subtract the third.
  - Thus
$$\widehat{\mathrm{V}}_{\text{two-way}}[\widehat{\boldsymbol{\beta}}] = \widehat{\mathrm{V}}_{G}[\widehat{\boldsymbol{\beta}}] + \widehat{\mathrm{V}}_{H}[\widehat{\boldsymbol{\beta}}] - \widehat{\mathrm{V}}_{G\cap H}[\widehat{\boldsymbol{\beta}}],$$
- Theory presented in Cameron, Gelbach, and Miller (2006, 2011), Miglioretti and Heagerty (2006), and Thompson (2006)
  - 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 $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ is not positive-definite (small G, H) then
  - Decompose $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = U\Lambda U'$; U contains eigenvectors of $\widehat{\mathrm{V}}$, and $\Lambda = \text{Diag}[\lambda_1, ..., \lambda_d]$ contains eigenvalues.
  - Create $\Lambda^+ = \text{Diag}[\lambda_1^+, ..., \lambda_d^+]$, with $\lambda_j^+ = \max\left(0, \lambda_j\right)$, and use $\widehat{\mathrm{V}}^+[\widehat{\boldsymbol{\beta}}] = U\Lambda^+ U'$
  - Stata add-on `cgmreg.ado`
  - 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 $G \longrightarrow \infty$ and $H \longrightarrow \infty$ then each fixed effect is estimated using many observations.
  - In practice the main consequence of including fixed effects is a reduction in within-cluster correlation of errors.

Hersch - Cross-Section with Two-Way Clustering

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

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

- Ideally cluster on both industry and occupation.

- 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)

		Variable			
		Industry Injury Rate		Occupation Injury Rate	
Estimated slope coefficient:		-1.894		-0.465	
Estimated standard errors and p-values:	Default (iid)	(0.415)	{0.0000}	(0.235)	{0.0478}
	Heteroscedastic robust	(0.397)	{0.0000}	(0.260)	{0.0737}
	One-way cluster on Industry	(0.643)	{0.0032}	(0.251)	{0.0639}
	One-way cluster on Occupation	(0.486)	{0.0001}	(0.363)	{0.2002}
	Two-way clustering	(0.702)	{0.0070}	(0.357)	{0.1927}

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 $\widehat{\mathbf{B}}$ in preceding has the form $\sum_i \sum_j w(i,j)\,\mathbf{x}_i\mathbf{x}_j'\widehat{u}_i\widehat{u}_j$.
  - Two-way clustering: $w(i,j) = 1$ for observations that share a cluster.
  - White and Domowitz (1984) time series: $w(i,j) = 1$ for observations "close" in time to one another.
  - Conley (1999) spatial: $w(i,j)$ decays to 0 as the distance between observations grows.
- The difference: White & Domowitz and Conley use mixing conditions to ensure decay of dependence in time or distance.
  - Mixing conditions do not apply to clustering due to common shocks.
  - Instead two-way robust requires independence across clusters.
- 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 $T \longrightarrow \infty$.
  - 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 $\Omega_g = \mathrm{E}[\mathbf{u}_g \mathbf{u}_g' | \mathbf{X}_g]$, such as within-cluster equicorrelation.
- Given a consistent estimate $\widehat{\Omega}$ of Ω, the feasible GLS estimator of $\boldsymbol{\beta}$ is

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

- To guard against misspecified Ω_g uses cluster-robust variance estimate

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

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

Examples of FGLS

- Random effects
  - One-way: $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g + \varepsilon_{ig}$ where α_g and ε_{ig} are i.i.d. errors.
  - Two-way: generalizes to $y_{igh} = \mathbf{x}_{igh}'\boldsymbol{\beta} + \alpha_g + \delta_h + \varepsilon_{ig}$ with i.i.d. errors
    - ★ but cannot then get cluster-robust variance matrix.
- Hierarchical linear models or mixed models: $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta}_g + u_{ig}$ and $\boldsymbol{\beta}_g = \mathbf{W}_g\boldsymbol{\gamma} + \mathbf{v}_j$ where u_{ig} and $\mathbf{v}_g$ are errors.
- Time series correlation for panel data
  - Kiefer (1980) $\Omega_g = \Omega$ and $\widehat{\Omega}_{ij} = G^{-1}\sum_{g=1}^{G}\widehat{u}_{ig}\widehat{u}_{jg}$, where $\widehat{u}_{ig}$ are OLS residuals
    - ★ Hausman and Kuersteiner (2008) bias correct for FE models.
  - AR(p) model for errors
    - ★ Hansen (2007b) bias corrects for FE models
- Spatial: Conley (1999) analyzes GMM with spatial correlation.

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 $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = m(\mathbf{x}_{ig}'\boldsymbol{\beta})$ 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 $G \longrightarrow \infty$.

- Can extend this approach to ML
  - specify a conditional density $f(y_{ig}|\mathbf{x}_{ig})$
  - do quasi-MLE with sandwich variance matrix estimate that controls for clustering
  - requires that conditional density $f(y_{ig}|\mathbf{x}_{ig})$ still correct when clustering is present.

Cluster-specific effects models

- Introduce cluster-specific effects α_g to control for within cluster correlation.
  - Parametric models: density specified for $f(y_{ig}|\mathbf{x}_{ig}, \boldsymbol{\beta}, \alpha_g)$, α_g not observed.
  - Conditional mean models: $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}, \alpha_g] = g(\mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g)$, α_g not observed.
- Fixed effects approach: α_g are parameters to be estimated
  - incidental parameters problem if asymptotics N_g is fixed while $G \longrightarrow \infty$
    - ★ there are N_g parameters $\alpha_1, ..., \alpha_G$ to estimate and $G \longrightarrow \infty$
    - ★ in general this contaminates estimation of $\boldsymbol{\beta}$ so that $\widehat{\boldsymbol{\beta}}$ is a inconsistent.
    - ★ notable exceptions are logit model, Poisson and nonlinear regression model with additive error.
- Random effects: α_g has density $h(\alpha_g|\boldsymbol{\eta})$
  - integrate out α_g
  - often no analytical solution but numerical methods work well as just a one-dimensional integral
  - results fragile to assumption about distribution of α_g (often $\mathcal{N}[0, \sigma_\alpha^2]$).

Instrumental variables

- Cluster-robust formula is easily adapted to instrumental variables estimation
  - In the just-identified case with $\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ use

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

    - ★ Shore-Sheppard (1996) showed the importance of doing so, extending Moulton to the IV setting.
- Hoxby and Paserman (1998) proposed a cluster-robust over-identifying restrictions test.
- 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`.

GMM

- Cluster-robust extends simply to GMM.
- For the g^{th} cluster the moment condition is
  - $\mathrm{E}[\mathbf{h}_g(\mathbf{w}_g, \theta)] = \mathbf{0}$ where $\mathbf{w}_g$ denotes all variables in the cluster
  - assume independence across clusters and $G \to \infty$.
- GMM estimator $\widehat{\theta}_{\text{GMM}}$ minimizes $\left(\sum_g \mathbf{h}_g\right)' \mathbf{W} \left(\sum_g \mathbf{h}_g\right)$, where $\mathbf{h}_g = \mathbf{h}_g(\mathbf{w}_g, \theta)$.
- The variance matrix estimate is

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

  - $\widehat{\mathbf{A}} = \sum_g \partial \mathbf{h}_g / \partial \theta' \big|_{\widehat{\theta}}$
  - $\widehat{\mathbf{B}} = \sum_g \widehat{\mathbf{h}}_g \widehat{\mathbf{h}}_g'$ 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 `vce(bootstrap, cluster(cid))`
  - asymptotically equivalent to use option `vce(cluster cid)`
  - equivalently can use prefix command `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 `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.

References

Acemoglu, D., and J.-S. Pischke (2003), "Minimum Wages and On-the-job Training," *Research in Labor Economics*, 22, 159-202.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Group Estimators," Oxford Bulletin of Economics and Statistics, 49, 431-434.

Bertrand, M., E. Duflo, and S. Mullainathan (2004), "How Much Should we Trust Differences-in-Differences Estimates," Quarterly Journal of Economics, 249-275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2009), "Inference with Dependent Data Using Cluster Covariance Estimators," manuscript, University of Chicago.

Bhattacharya, D. (2005), "Asymptotic inference from multi-stage samples,"' Journal of Econometrics, 126: 145-171.

Cameron, A. C., Gelbach, J. G., and D. L. Miller (2006, 2011). "Robust Inference with Multi-Way Clustering." NBER Technical Working Paper 0327 and J. Business and Economic Statistics, forthcoming.

Cameron, A. C., Gelbach, J. G., and D. L. Miller (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," Review of Economics and Statistics, 90, 414-427.

Cameron, A.C., and N. Golotvina (2005), "Estimation of Country-Pair Data Models Controlling for Clustered Errors: with International Trade Applications," U.C.-Davis Economics Department Working Paper No. 06-13.

Cameron, A.C., and D. L. Miller (2011), "Robust Inference with Clustered Data", in A. Ullah and D. E. Giles eds., Handbook of Empirical Economics and Finance, CRC Press, pp.1-28.

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

Conley, T. G. (1999), "GMM with cross sectional dependence," Journal of Econometrics, 92: 1-45.

Donald, S. and K. Lang (2004), "Inferences with Differences in Differences and Other Panel Data," October 22, 2004.

Driscoll, J. C., and A.C. Kraay (1998), "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," Review of Economics and Statistics, 80: 549-560.

Fafchamps, M., and F. Gubert (2006), "The Formation of Risk Sharing Networks," mimeo, April 2006.

Finlay, K., and L. M. Magnusson (2009), "Implementing Weak Instrument Robust Tests for a General Class of Instrumental-Variables Models," Stata Journal, 9: 398-421.

Foote, C. L. (2007), "Space and Time in Macroeconomic Panel Data: Young Workers and State-Level Unemployment Revisited," Working Paper 07-10, Federal Reserve Bank of Boston.

Hansen, C. B. (2007a), "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," Journal of Econometrics, 140, 670-694.

Hansen, C. B. (2007b), "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," Journal of Econometrics, 141: 597-620.

Hersch, J. (1998), "Compensating Wage Differentials for Gender-Specific Job Injury Rates," American Economic Review, 88: 598-607.

Hoxby, C., and M. D. Paserman (1998), "Overidentification Tests with Group Data," NBER Technical Working Paper 0223.

Ibragimov, R. and U.K. Muller (2007), "T-Statistic Based Correlation and Heterogeneity Robust Inference," Harvard Institute of Economic Research Discussion Paper No. 2129.

Kézdi, G. (2004), "Robust Standard Error Estimation in Fixed-Effects Models. Robust Standard Error Estimation in Fixed-Effects Panel Models," Hungarian Statistical Review, Special Number 9: 95-116.

Kloek, T. (1981), "OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated," Econometrica, 49: 205-07.

Kuersteiner, G. and J. Hausman (2007), "Difference in Difference meets Generalized Least Squares: Higher Order Properties of Hypotheses Tests".

Liang, K.-Y., and S.L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," Biometrika, 73, 13-22.

Miglioretti, D.L., and P.J. Heagerty (2006), "Marginal Modeling of Nonnested Multilevel Data using Standard Software," *American Journal of Epidemiology*, 165(4), 453-463.

Moulton, B.R. (1986), "Random Group Effects and the Precision of Regression Estimates," Journal of Econometrics, 32, 385-97.

Moulton, B.R. (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," Review of Economics and Statistics, 72, 334-38.

Pepper, J. V. (2002), "Robust Inferences from Random Clustered Samples: An Application using Data from the Panel Study of Income Dynamics," Economics Letters, 75: 341-5.

Petersen, M. (2009), "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," Review of Financial Studies, 22, 435-480.

Rogers, W.H. (1993), "Regression Standard Errors in Clustered Samples," Stata Technical Bulletin, 13, 19-23.

Scott, A. J., and D. Holt (1982), "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods," J. Am. Stat. Assoc., 77: 848-854.

Shore-Sheppard, L. (1996), "The Precision of Instrumental Variables Estimates with Grouped Data," Working Paper 374, Princeton University Industrial Relations Section.

Thompson, S. (2005), "A Simple Formula for Standard Errors that Cluster by Both Firm and Time," unpublished manuscript.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 48: 817-838.

White, H. (1984), Asymptotic Theory for Econometricians, San Diego, Academic Press.

White, H., and I. Domowitz (1984), "Nonlinear Regression with Dependent Observations," Econometrica, 52: 143-162.

Wooldridge, J.M. (2003), "Cluster-Sample Methods in Applied Econometrics," American Economic Review, 93, 133-138.

Wooldridge, J. M. (2006), "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis," unpublished manuscript, Michigan State University Department of Economics.