Day 2B
Inference for Clustered Data: Part 2
With a Cross-section Data Example
(Revised August 25 to add 14. Dyadic Clustering)

A. Colin Cameron
Univ. of Calif. - Davis

... for

Center of Labor Economics
Norwegian School of Economics
Advanced Microeconometrics

Aug 28 - Sep 1, 2017

1. Introduction

- Consider clustering in more detail.
- The example is a Moulton-type data example
  - data on individuals in households in communes (villages)
- These slides are a summary of
  - A. Colin Cameron and Douglas L. Miller (2015), "A Practitioner's Guide to Robust Inference with Clustered Data," *Journal of Human Resources*, Vol.50 (2, Spring), 317-373.
  - Preprint version available at
    - http://cameron.econ.ucdavis.edu/research/papers.html

Outline

2. Moulton-Type Data

- Vietnam data on individuals in households in communes
  - pharvis - number of direct pharmacy visits in past 12 months
  - lnhhexp - log of household medical expenditures
  - illness - number of illnesses
- Identifiers
  - commune - identifies the commune
  - hh - created variable that identifies household
  - person_in_hh - created variable that identifies person in household
  - person - created variable that uniquely identifies household
- We analyze using (1) regress

OLS without using Panel Commands

- Summary statistics

```
. sum pharvis lnhhexp illness AGE hh person_in_hh person
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pharvis	27765	.5117594	1.313427	0	30
lnhhexp	27765	2.60261	.6244145	.0467014	5.405502
illness	27765	.6219701	.8995068	0	9
AGE	27765	2.977504	.9671446	0	4.59512
hh	27765	3098.336	1601.742	1	5740
person_in_hh	27765	3.296957	1.97824	1	19
person	27765	309836.9	160174.5	101	574004

Cluster-Robust Variance for OLS

- * OLS estimation with cluster-robust standard errors

```
* Cluster on household and then on commune
quietly regress pharvis lnhhexp illness
estimates store OLS_iid
quietly regress pharvis lnhhexp illness, vce(robust)
estimates store OLS_het
quietly regress pharvis lnhhexp illness, vce(cluster hh)
estimates store OLS_hh
quietly regress pharvis lnhhexp illness, vce(cluster commune)
estimates store OLS_comm
estimates table OLS_iid OLS_het OLS_hh OLS_comm, b(%10.4f) se stats(r2 N)
```

- Standard errors increase with broader clustering level

```
. estimates table OLS_iid OLS_het OLS_hh OLS_comm, b(%10.4f) se stats(r2 N)
```

Variable	OLS_iid	OLS_het	OLS_hh	OLS_comm
lnhhexp	0.0248	0.0248	0.0248	0.0248
	0.0115	0.0109	0.0140	0.0211
illness	0.6241	0.6241	0.6241	0.6241
	0.0080	0.0141	0.0183	0.0342
_cons	0.0591	0.0591	0.0591	0.0591
	0.0316	0.0292	0.0367	0.0556
r2	0.1818	0.1818	0.1818	0.1818
N	27764	27764	27764	27764

legend: b/se

3. Analysis using Panel Commands

- Suppose we cluster on household and want to use xtdescribe

```
. xtset hh
       panel variable:  hh (unbalanced)

. xtdescribe
must specify timevar; use xtset
r(459);

. xtset hh person_in_hh
       panel variable:  hh (unbalanced)
        time variable:  person_in_hh, 1 to 19
                delta:  1 unit
```

- Also need to give a "time" identifier - here person_in_hh

```
. xtdescribe

      hh:  1, 2, ..., 5740                                  n =       5740
person_in_hh:  1, 2, ..., 19                                T =         19
           Delta(person_in_hh) = 1 unit
           Span(person_in_hh)  = 19 periods
           (hh*person_in_hh uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       2       4         5         6       8      19

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------------------
     1376     23.97   23.97 |  1111...............
     1285     22.39   46.36 |  11111..............
      853     14.86   61.22 |  111111.............
      706     12.30   73.52 |  111................
      471      8.21   81.72 |  1111111............
      441      7.68   89.41 |  11.................
      249      4.34   93.75 |  11111111...........
      126      2.20   95.94 |  1..................
      125      2.18   98.12 |  111111111..........
      108      1.88  100.00 | (other patterns)
 ---------------------------+---------------------
     5740    100.00         |  XXXXXXXXXXXXXXXXXXX
```

Intracluster Correlation

- Cluster here is household - here fairly high intracluster correlation

```
. loneway pharvis hh

                One-way Analysis of Variance for pharvis:

                                        Number of obs =      27765
                                            R-squared =     0.3878

    Source                    SS        df       MS            F     Prob > F
-------------------------------------------------------------------------------
Between hh                 18571.572   5739    3.2360293      2.43     0.0000
Within hh                  29323.838  22025    1.3313888
-------------------------------------------------------------------------------
Total                      47895.411  27764    1.7250904

         Intraclass        Asy.
         correlation       S.E.        [95% Conf. Interval]
         ------------------------------------------------
            0.22825      0.00641        0.21569     0.24081

         Estimated SD of hh effect                .6275084
         Estimated SD within hh                   1.153858
         Est. reliability of a hh mean             0.58857
              (evaluated at n=4.84)
```

- Cluster here is household
  - other methods to estimate intraclass correlation

```
. quietly xtreg pharvis, mle

.
. display "Intra-class correlation for household: " e(rho)
Intra-class correlation for household: .22283723

.
. quietly correlate pharvis L1.pharvis

.
. display "Correlation for adjoining household:   " r(rho)
Correlation for adjoining household:   .20441495
```

```
* OLS, RE and FE estimation with clustering on household and on village
quietly regress pharvis lnhhexp illness, vce(cluster hh)
estimates store OLS_hh
quietly xtreg pharvis lnhhexp illness, re
estimates store RE_hh
quietly xtreg pharvis lnhhexp illness, fe
estimates store FE_hh
quietly xtset commune
quietly regress pharvis lnhhexp illness, vce(cluster commune)
estimates store OLS_vill
quietly xtreg pharvis lnhhexp illness, re
estimates store RE_vill
quietly xtreg pharvis lnhhexp illness, fe
estimates store FE_vill
estimates table OLS_hh RE_hh FE_hh OLS_vill RE_vill FE_vill, b(%7.5f) se(%7.4f)
```

- Here `xt` commands cluster on household
  - but second lot of se's cluster on village (commune)
  - RE and FE more efficient than OLS
  - FE similar efficiency to RE as much within variation
    - to see this `xtsum pharvis`

```
. estimates table OLS_hh RE_hh FE_hh OLS_vill RE_vill FE_vill, b(%7.5f) se(%7.4f)
```

Variable	OLS_hh	RE_hh	FE_hh	OLS_vill	RE_vill	FE_vill
lnhhexp	0.02477	0.01839	(omitted)	0.02477	-0.04489	-0.06570
	0.0140	0.0168		0.0211	0.0149	0.0158
illness	0.62416	0.61713	0.60969	0.62416	0.61551	0.61407
	0.0183	0.0083	0.0096	0.0342	0.0081	0.0082
_cons	0.05909	0.08549	0.13255	0.05909	0.24306	0.30082
	0.0367	0.0448	0.0087	0.0556	0.0441	0.0426

legend: b/se

4. Mixed or Multi-level or Hierarchical Model

- Not used in microeconometrics but used in many other disciplines.
- Stack all observations for cluster g and specify

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{Z}_g \mathbf{u}_g + \boldsymbol{\varepsilon}_g$$

  where $\mathbf{u}_g$ is iid $(\mathbf{0}, \mathbf{G})$ and $\mathbf{Z}_g$ is called a design matrix and $\boldsymbol{\varepsilon}_g \sim (\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$.
- Random effects: $\mathbf{Z}_g = \mathbf{e}$ (a vector of ones) and $\mathbf{u}_g = \alpha_g$
- Random coefficients: $\mathbf{Z}_g = \mathbf{X}_g$
  - Reason: $\boldsymbol{\beta}_g \sim (\boldsymbol{\beta}, \Sigma)$ so $\boldsymbol{\beta}_g \sim \boldsymbol{\beta} + \mathbf{u}_g$ where $\mathbf{u}_g \sim (\mathbf{0}, \Sigma)$
  - So $\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_g = \mathbf{X}_g (\boldsymbol{\beta} + \mathbf{u}_g) + \boldsymbol{\varepsilon}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{X}_g \mathbf{u}_i + \boldsymbol{\varepsilon}_g$.
  - Note: $\mathbf{y}_g | \mathbf{X}_g$ has mean $\mathbf{X}_g \boldsymbol{\beta}$ and variance $\mathbf{X}_g' \Sigma \mathbf{X}_g + \sigma_\epsilon^2 \mathbf{I}$.
- Simplest case of random intercept is same as `xtreg, mle`

```
. mixed lwage exp exp2 wks ed || hh:
```

- Example where `illness` has random slope

```
. * Mixed model with random interecept and a random slope
. mixed pharvis lnhhexp illness || hh: illness, nolog covariance(unstructured)

Mixed-effects ML regression                     Number of obs     =     27,765
Group variable: hh                              Number of groups  =      5,740

                                                Obs per group:
                                                              min =          1
                                                              avg =        4.8
                                                              max =         19

                                                Wald chi2(2)      =    2376.50
Log likelihood = -38849.338                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     pharvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnhhexp |   .0035325   .0098986     0.36   0.721    -.0158684    .0229334
     illness |   .7688076   .0157887    48.69   0.000     .7378624    .7997529
       _cons |   .0652403   .0269342     2.42   0.015     .0124503    .1180303
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
hh: Unstructured             |
                var(illness) |   .8113512   .0228817      .7677208    .8574612
                  var(_cons) |   .0001267   .0000131      .0001034    .0001553
          cov(illness,_cons) |   .0101402   .0005424      .0090772    .0112033
-----------------------------+------------------------------------------------
               var(Residual) |   .7100916    .006757      .6969708    .7234594
------------------------------------------------------------------------------
LR test vs. linear model: chi2(3) = 10661.04              Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Extensions

- Proceeding example was a two-level model.
- Can add additional levels

e.g. people in households in communes

- And can have two-way random effects
  - $y_{igh} = \alpha_g + \delta_g + \mathbf{x}'_{igh}\boldsymbol{\beta} + \varepsilon_{igh}$
  - α_g iid $(\alpha, \sigma^2_\alpha)$ and δ_h iid $(0, \sigma^2_\delta)$.
- Code as
  - ```
    * Twoway random effects with error
    *       e_g + e_h + e_igh, g=ILLDAYS h=hh
    ```
  - ```
    mixed pharvis lnhhexp illness || _all:  R.ILLDAYS || hh:
    , mle
    ```

5. Cluster-Specific Fixed Effects Model

- The model is $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g + u_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + \sum_{h=1}^{G} \alpha_g dh_{ig} + u_{ig}$
  - there are G dummy variables $d1_{ig}, ..., dG_{ig}$
  - $dh_{ig} = 1$ if ig^{th} observation is in cluster h and $= 0$ otherwise.
- Do FE's Eliminate Within-Cluster Error Correlation? No.
- Does CRVE still work? Yes if $G \rightarrow \infty$ and either N_g fixed or $N_g \rightarrow \infty$.
- What is rank of CRVE with K regressors and G clusters
  - For OLS it is minimum of K and $G - 1$
    - ★ Reason: $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{B}}\ (\mathbf{X}'\mathbf{X})^{-1}$
    - ★ $\widehat{\mathbf{B}} = \mathbf{C}'\mathbf{C}$, where $\mathbf{C}' = [\mathbf{X}_1'\widehat{\mathbf{u}}_1 \cdots \mathbf{X}_G'\widehat{\mathbf{u}}_G]$ is $K \times G$
    - ★ and $\mathbf{X}_1'\widehat{\mathbf{u}}_1 + \cdots + \mathbf{X}_G'\widehat{\mathbf{u}}_G = \mathbf{0}$
- For FE it is also minimum of K and $G - 1$

Feasible GLS (with FE's)

- More difficult with fixed effects if N_g is small.
- If N_g is finite then $\widehat{\alpha}_g$ is inconsistent for a_g
  - Does not lead to inconsistent $\widehat{\boldsymbol{\beta}}$
  - But does mean residuals $\widehat{u}_{ig}$ inconsistently estimated
  - This contaminates $\widehat{\Omega}$ used in FGLS estimation.
- Hansen (2007b) provides bias-corrected FGLS for AR(p) errors
  - Brewer, Crossley and Joyce (2013) implement in DiD setting and show power gains
- Hausman and Kuersteiner (2008) provide bias-corrected FGLS for Kiefer (1980) error model
  - $\Omega_g = \Omega$ and $\widehat{\Omega}_{ij} = G^{-1} \sum_{g=1}^{G} \widehat{u}_{ig} \widehat{u}_{jg}$, where $\widehat{u}_{ig}$ are OLS residuals

Testing the Need for Fixed Effects

- Hausman test: $T_{\mathrm{H}} = (\widehat{\boldsymbol{\beta}}_{1;\mathrm{FE}} - \widehat{\boldsymbol{\beta}}_{1;\mathrm{RE}})'\widehat{\mathrm{V}}^{-1}(\widehat{\boldsymbol{\beta}}_{1;\mathrm{FE}} - \widehat{\boldsymbol{\beta}}_{;\mathrm{RE}})$,
- Must use a modified Hausman test
  - Reason: Default Hausman uses $\widehat{\mathrm{V}} = \widehat{\mathrm{V}}_{1;\mathrm{FE}} - \widehat{\mathrm{V}}_{1;\mathrm{RE}}$
  - But this requires that RE model is fully efficient under H_0
- Wooldridge (2012, p.332): OLS regression

$$y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + \overline{\mathbf{w}}_g'\boldsymbol{\gamma} + u_{ig},$$

  - where $\mathbf{w}_g$ denotes the subcomponent of $\mathbf{x}_{ig}$ that varies within cluster and $\overline{\mathbf{w}}_g = N_g^{-1}\sum_{i=1}^{G}\mathbf{w}_{ig}$.
  - test $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ using a Wald test based on a CRVE
  - FE model is necessary if we reject H_0
  - Stata user-written command `xtoverid`, due to Hoechle (2007).
- Or do pairs cluster bootstrap to estimate $\widehat{\mathrm{V}}$.

6. 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.

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 $\mathbf{x}$ handles weighting
- And stratified
  - this improves estimator efficiency somewhat
- Bhattacharya (2005) gives a general GMM treatment.

- Econometricians reasonably
  - Cluster on PSU or higher
  - Sometimes weight and sometimes not
  - 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]`

7. Multi-way Clustering

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

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

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

Two-way clustering

- Robust variance matrix estimates are of the form

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

- 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 $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\beta}$ 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

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

- Theory presented in Cameron, Gelbach, and Miller (2006, 2011), Miglioretti and Heagerty (2006), and Thompson (2006, 2011)
  - Extends to multi-way clustering.
- Early empirical applications that independently proposed this method include Acemoglu and Pischke (2003).

Implementation

- If $\widehat{V}[\widehat{\beta}]$ is not positive-definite (small G, H) then
  - Decompose $\widehat{V}[\widehat{\beta}] = U\Lambda U'$; U contains eigenvectors of $\widehat{V}$, 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 $\widehat{V}^+[\widehat{\beta}] = U\Lambda^+ U'$
  - Stata add-on `cgmreg.ado` implements this.
- Also Stata add-on `ivreg2.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.

Cluster on Household and Age

- Use cgmreg.ado
  - at cameron.econ.ucdavis.edu/research/papers.html

```
. cgmreg pharvis lnhhexp illness, cluster(hh AGE)

Note: +/- means the corresponding matrix is added/subtracted

Calculating cov part for variables:  hh (+)
Calculating cov part for variables:  hh AGE (-)
Calculating cov part for variables:  AGE (+)

                                                Number of obs      =    27765
                                                Num clusvars       =    2
                                                Num combinations   =    3

                                                G(hh)              =    5740
                                                G(AGE)             =    98
```

pharvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnhhexp	.0247682	.0139228	1.78	0.075	-.00252	.0520563
illness	.624163	.0190994	32.68	0.000	.5867288	.6615971
_cons	.0590868	.0382208	1.55	0.122	-.0158245	.1339981

- User-written ivreg2 command gives similar (different dof)

```
. ivreg2 pharvis lnhhexp illness, cluster(hh AGE)

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on hh and AGE

Number of clusters (hh) =          5740              Number of obs =     27765
Number of clusters (AGE) =           98              F(  2,     97) =    534.02
                                                     Prob > F      =    0.0000
Total (centered) SS     =   47895.41055              Centered R2   =    0.1818
Total (uncentered) SS   =         55167              Uncentered R2 =    0.2897
Residual SS             =   39185.73296              Root MSE      =     1.188

------------------------------------------------------------------------------
             |               Robust
     pharvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnhhexp |   .0247682    .013877     1.78   0.074    -.0024302    .0519666
     illness |    .624163   .0190356    32.79   0.000     .5868539    .6614721
       _cons |   .0590868   .0380856     1.55   0.121    -.0155595    .1337331
------------------------------------------------------------------------------
Included instruments: lnhhexp illness
------------------------------------------------------------------------------
```

Application

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

Feasible GLS

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

Spatial Correlation

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

Spatial Correlation Consistent VE

- Driscoll and Kraay (1998) panel data when $T \to \infty$
  - 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.
- Barrios, Diamond, Imbens, and Kolesár (2012) state-year panel on individuals with spatial correlation across states. And use randomization inference.

8. Few Clusters: Inference with few clusters

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

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

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

The Basic Problem with Few Clusters

- OLS overfits with $\widehat{u}$ systematically biased to zero compared to u.
  - e.g. OLS with iid normal errors $\mathrm{E}[\widehat{\mathbf{u}}'\widehat{\mathbf{u}}] = (N-K)\sigma^2$, not $N\sigma^2$.
- Problem is greatest as G gets small - "few" clusters.
- How few is few?
  - balanced data; $G < 20$ to $G < 50$ depending on data
  - unbalanced data: G less than this.
- Unusual situation for applied econometrics
  - since have many observations estimation is reasonably precise
  - so it is worthwhile doing statistical inference
  - but because G is small the usual asymptotic theory leads to invalid inference.

Solutions

- 1. Bias-corrected CRVE
- 2. Cluster-Bootstrap with Asymptotic Refinement
- 3. Improved Critical Values
- This is an active area of research
  - the following discussion needs references updated and newer ones added.

9. Bias-Corrected CRVE

- Simplest is $\widetilde{\mathbf{u}}_g = \sqrt{c}\widehat{\mathbf{u}}_g$, already mentioned.
- CR2VE generalizes HC2 for heteroskedasticity
  - $\widetilde{\mathbf{u}}_g^* = [\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1/2}\widehat{\mathbf{u}}_g$ where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$
  - gives unbiased CRVE if errors iid normal
- CR3VE generalizes HC3 for heteroskedasticity
  - $\widetilde{\mathbf{u}}_g^+ = \sqrt{G/(G-1)}[\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g$ where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$
  - same as jackknife
- Problems:
  - Not clear which is better CR2VE or CR3VE
  - More importantly, still over-rejects when few clusters.

10. Few Clusters: Cluster bootstrap with asymptotic refinement

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

Stata Pairs Cluster Bootstrap BC and BCa Confidence Intervals

- Keep only 20 communes (Cluster on commune)
  - Stata gives bias-corrected and accelerated BC intervals

```
preserve
keep if commune <= 20
(25122 observations deleted)
xtset commune // for cluster bootstrap can only xtset the cluster variable
regress pharvis lnhhexp illness, vce(boot, ///
cluster(commune) seed(10101) reps(999) bca)
(running regress on estimation sample)
Jackknife replications (20)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
....................
Bootstrap replications (999)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................. 50
```

```
. estat bootstrap, all

Linear regression                               Number of obs     =      2,643
                                                Replications      =        999

                                (Replications based on 20 clusters in commune)

------------------------------------------------------------------------------
             |    Observed               Bootstrap
     pharvis |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnhhexp |  -.14339762  -.0026205   .05623308   -.2536124  -.0331828   (N)
             |                                      -.2590029  -.0388589   (P)
             |                                      -.2551242  -.0388589  (BC)
             |                                      -.2517517  -.0357437 (BCa)
     illness |   .63346949   .0080904   .07694949    .4826513   .7842877   (N)
             |                                       .5227506   .8170654   (P)
             |                                        .524489    .825703  (BC)
             |                                       .5255618   .8272487 (BCa)
       _cons |   .65234094   .0070333   .20171892    .2569791   1.047703   (N)
             |                                       .2862576   1.060108   (P)
             |                                       .2862576   1.059884  (BC)
             |                                       .2717655   1.055986 (BCa)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval
```

Stata Pairs Cluster Bootstrap Percentile-t Confidence Intervals

- In theory the BC and BCa confidence intervals provide asymptotic refinement.
  - but we find they differ little from the percentile bootstrap.
- another possibility is to Resample cluster pairs as for cluster robust se's bootstrap
- But at each bootstrap calculate $w_b^* = (\widehat{\beta}_b^* - \widehat{\beta})/s_{\widehat{\beta}_b^*}$
  - Note: we subtract $\widehat{\beta}$ because the bootstrap views the population as the sample so the dgp value of β is $\widehat{\beta}$

```
. * Percentile-t for a single coefficient: Bootstrap the t statistic
. quietly regress pharvis lnhhexp illness, vce(cluster commune)
. local theta = _b[lnhhexp]
. local setheta = _se[lnhhexp]
. bootstrap tstar=((_b[lnhhexp]-`theta')/_se[lnhhexp]), seed(10101) ///
> reps(999) saving(percentilet, replace): ///
> regress pharvis lnhhexp illness, vce(cluster commune)
```

- We have 999 values of w_b^*, called `tstar` below.

```
. * Percentile-t p-value for symmetric two-sided Wald test of H0: theta = 0
. use percentilet, clear
(bootstrap: regress)

. quietly count if abs(`theta'/`setheta') < abs(tstar)

. display "p-value = " r(N)/_N
p-value = .00500501

.
. * Percentile-t critical values and confidence interval
. _pctile tstar, p(2.5,97.5)

. scalar lb = `theta' + r(r1)*`setheta'

. scalar ub = `theta' + r(r2)*`setheta'

. display "2.5 and 97.5 percentiles of t* distn: " r(r1) ", " r(r2) _n ///
>     "95 percent percentile-t confidence interval is:  (" lb   ","  ub ")"
2.5 and 97.5 percentiles of t* distn: -1.7114737, 1.5089508
95 percent percentile-t confidence interval is:  (-.23920977,-.05892316)
```

Wild Cluster Bootstrap

1. Obtain the OLS estimator $\widehat{\beta}$ 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 $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g' \widehat{\beta} + \widehat{\mathbf{u}}_g^*$.
      This yields wild cluster bootstrap resample $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$.
   2. Calculate the OLS estimate $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ and given these form the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \widehat{\beta}_1)/s_{\widehat{\beta}_{1,b}^*}$.
3. Reject H_0 at level α if and only if

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

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

Current Research

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

Use the Bootstrap with Caution

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

11. Few Clusters: Improved T Critical Values

- Suppose all regressors are invariant within clusters, clusters are balanced and errors are i.i.d. normal
  - then $y_{ig} = \mathbf{x}'_g \boldsymbol{\beta} + \varepsilon_{ig} \Longrightarrow \bar{y}_g = \bar{\mathbf{x}}'_g \boldsymbol{\beta} + \bar{\varepsilon}_g$ with $\bar{\varepsilon}_g$ i.i.d. normal
  - so Wald test based on OLS is exactly $T(G-L)$, where L is the number of group invariant regressors.
- Extend to nonnormal errors and group varying regressors
  - asymptotic theory when G is small and $N_g \to \infty$.
  - 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.

- Imbens and Kolesar (2012)
  - Data-determined number of degrees of freedom for t and F tests
  - Builds on Satterthwaite (1946) and Bell and McCaffrey (2002).
  - Assumes normal errors and particular model for Ω.
  - Match first two moments of test statistic with first two moments of χ^2.
  - $v^* = (\sum_{j=1}^{G} \lambda_j)^2 / (\sum_{j=1}^{G} \lambda_j^2)$ and λ_j are the eigenvalues of the $G \times G$ matrix $\mathbf{G}'\widehat{\Omega}\mathbf{G}$.
  - Find works better than 2-point Wild cluster bootstrap but they did not impose H_0.

- Carter, Schnepel and Steigerwald (2013)
  - provide asymptotic theory when clusters are unbalanced
  - propose a measure of the effective number of clusters
  - $G^* = G/(1+\delta)$
    - where $\delta = \frac{1}{G}\sum_{g=1}^{G}\{(\gamma_g - \bar{\gamma})^2/\bar{\gamma}^2\}$
    - $\gamma_g = \mathbf{e}_k' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}_g'\Omega_g\mathbf{X}_g \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{e}_k$
    - $\mathbf{e}_k$ is a $K \times 1$ vector of zeroes aside from 1 in the k^{th} position if $\widehat{\beta} = \widehat{\beta}_k$
    - $\bar{\gamma} = \frac{1}{G}\sum_{g=1}^{G}\gamma_g$.
- Cluster heterogeneity $(\delta \neq 0)$ can arise for many reasons
  - variation in N_g, variation in $\mathbf{X}_g$ and variation in ˙$_g$ across clusters.
- Brewer, Crossley and Joyce (2013)
  - Do FGLS as gives both efficiency gains and works well even with few clusters.

12. Few Clusters: Special Cases

- Bester, Conley and Hansen (2009)
  - obtain $T(G-1)$ in settings such as panel where mixing conditions apply.
- Ibragimov and Muller (2010) take an alternative approach
  - suppose only within-group variation is relevant
  - then separately estimate $\widehat{\boldsymbol{\beta}}_g s$ and average
  - asymptotic theory when G is small and $N_g \rightarrow \infty$
- 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 $y_{st} = \mathbf{x}_{st}'\boldsymbol{\beta} + \mathbf{z}_t'\gamma + \varepsilon_{ig}$ where $\mathbf{z}_t$ are for example time dummies.
- This limitation is relevant in DiD models with few treated groups
  - Conley and Taber (2010) present a novel method for that case.
  - for synthetic control methods (one treated group) see Abadie, Diamond and Hainmueller (2010)
    - inference not established.

13. Extensions

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

Extensions (continued)

- See Cameron and Miller, JHR (2015) paper.
- 7.1 Cluster-Robust F-tests
- 7.2 Instrumental Variables Estimators
  - IV, 2SLS, linear GMM
  - Need modified Hausman test for endogeneity : `estat endogenous`
  - Weak instruments:
    - First-stage F-test should be cluster-robust
    - use add-on `xtivreg2`
    - Finlay and Magnusson (2009) have Stata add-on `rivtest.ado`.
- 7.3 Nonlinear Estimators
  - Population-averaged (`xtreg, pa`) and random effects (e.g. `xtlogit, re`) give quite different βs
  - Rarely can eliminate fixed effects if N_g is small.
- 7.4 Cluster-randomized Experiments

14 ADDENDUM: Dyadic Clustering

- A dyad is a pair. An example is country pairs.
- Gravity model for volume of trade between countries
  - Does WTO/GATT membership increase trade? (Rose AER 2004)
  - Data on 187 countries and 52 years (1948-1999)
  - $N = 234{,}597$
  - $\leq$ max observations $= 52 \times 178 \times 177/2 = 819{,}156$.
- Model estimated by OLS is

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

  where y_{ght} is the natural logarithm of real bilateral trade trade between countries g and h in time period t.
- How do we compute standard errors controlling for correlation in u_{ght}?
  - Complication is that US-UK error likely correlated with any pair involving the US or involving UK.

- In general robust variance estimates are of form

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

- For two-way clustering with $y_{igh} = \mathbf{x}'_{igh}\beta + u_{igh}g$

$$\text{E}[u_{igh}u_{jg'h'}|x_{igh}, x_{jg'h'}] = 0, \text{ unless } g = g' \text{ or } h = h'.$$

  - use

$$\widehat{\mathbf{B}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}[g = g' \text{ and/or } h = h'] \times \widehat{u}_{igh}\widehat{u}_{jg'h'}\mathbf{x}_{igh}\mathbf{x}'_{jg'h'}.$$

- For dyadic we have

$$\text{E}[u_{igh}u_{jg'h'}|x_{igh}, x_{jg'h'}] = 0, \text{ unless } g = g' \text{ or } h = h' \text{ or } h = g' \text{ or } g = h'.$$

  - use

$$\widehat{\mathbf{B}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}[g = g' \text{ and/or } h = h' \text{ and/or } h = g' \text{ and/or } g = h'] \times \widehat{u}_{igh}\widehat{u}_{jg'h'}\mathbf{x}_{igh}\mathbf{x}'_{jg'h'}.$$

Example: G=4 countries

- Six Pairs $(1,2)$, $(1,3)$, $(1,4)$, $(2,3)$, $(2,4)$ and $(3,4)$
  - country-pair: only $(g,h)=(g',h')$ diagonal entries denoted CP
  - two-way: $g=g'$ and/or $h=h'$ denoted CP and 2way.
  - dyadic: also $g=h'$ or $h=g'$ denoted CP, 2way and DYAD.

(g,h)\(g′,h′)	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
(1,2)	CP	2way	2way	**DYAD**	**DYAD**	
(1,3)	2way	CP	2way	2way		**DYAD**
(1,4)	2way	2way	CP		2way	2way
(2,3)	**DYAD**	2way		CP	2way	**DYAD**
(2,4)	**DYAD**		2way	2way	CP	2way
(3,4)		**DYAD**	2way	**DYAD**	2way	CP

- For small G large fraction of correlation matrix is nonzero
  - $G=10$: 38% of error correlations are nonzero
  - $G=30$: 13% of error correlations are nonzero.
- For large G the fraction potentially correlated $\rightarrow 4/(G-1)$.

- It makes a big difference, even with panel data and country-pair FEs.
  - Table has standard error ratios - bold is dyadic versus usual method.

TABLE 4B: COUNTRY-PAIR PANEL DATA EXAMPLE WITH COUNTRY-PAIR FIXED EFFECTS

Comparison of Standard Errors computed in various ways

				SE RATIOS			
				DYAD	**DYAD**	DYAD	DYAD
	COEFF	TIME	SIG 5%	/HET	**/PAIRS**	/2WAY	/JACK
Both_in_GATTorWTO	0.1271	Yes	No	6.96	**3.16**	1.21	1.31
One_in_GATTorWTO	0.0600	Yes	No	5.62	**2.55**	1.14	1.28
GSP	0.1754	Yes	Yes	7.89	**2.99**	1.01	1.31
Log_Distance	--		--	--	**--**	--	--
Log_product_real_GDP	0.4425	Yes	Yes	8.69	**3.25**	1.11	1.32
Log_product_real_GDP_pc	0.2368	Yes	No	8.57	**3.15**	1.11	1.31
Regional_FTA	0.7639	Yes	Yes	6.76	**2.34**	1.09	1.24
Currency_Union	0.6314	Yes	Yes	3.51	**1.41**	1.26	1.07
Common_language	--		--	--	**--**	--	--
Land_border	--		--	--	**--**	--	--
Number_landlocked	--		--	--	**--**	--	--
Number_islands	--		--	--	**--**	--	--
Log_product_land_area	--		--	--	**--**	--	--
Common_colonizer	--		--	--	**--**	--	--
Currently_colonized	0.2957	Yes	Yes	2.22	**0.64**	1.24	0.70
Ever colony	--		--	--	**--**	--	
Common_country	--		--	--	**--**	--	--
Constant	0.0000		--	--	**--**	--	--
Year dummies	Yes						
Country-pair dummies	Yes						
Observations	234597						
R-squared	0.853						

Resources

- A. Colin Cameron and Douglas L. Miller, "Robust Inference for Dyadic Data", December 31, 2014.
  - Paper presented at the January 2015 ASSA meetings.
- Code for Dyadic Cluster Robust Standard Errors
  - For OLS our Stata code is regdyad2.ado
  - This is very much a "beta version". We are in the midst of editing and tweaking it, but we believe that it runs and works.
  - The syntax to use is
    - ★ regdyad2 yvar xvars, dyads(ID_var_1 ID_var_2)
- Paper and code downloadable at
  - http://cameron.econ.ucdavis.edu/research/papers.html

14. 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.