Lecture 18

Clustered Errors

Example 1: Individuals in clusters (Monte Carlo 1986, 1990) pre cluster robust

Example 2: Panel data
Different-in-differences state-year panel

Policy indicator $d_{ts}$ varies by state $(s)$ over time $(t)$

E.g. $d_{ts} = 1$ if minimum wage law is effect

OLS $y_{ts} = \alpha + \alpha_{ts} + \beta + \delta d_{ts} + \eta_{ts}$

$d_{ts}$ is highly correlated over time for given $s$
and $\eta_{ts}$ is not correlated over time for given $s$
Again, default OLS standard errors are way too small. The same problem arises if we have individual-level data:

$$y_{its} = \alpha + x_{its} \beta + \delta d_{ts} + \gamma_{its}$$

Bertrand, Duflo & Mullainathan (2004) found either people did not cluster-robust, or they did but erroneously on state-year pair.

- Need to cluster-robust on state.
**Intuition**

Univariate: \( y_i \sim (\mu, \sigma^2) \)

We want \( \text{Var}(\bar{y}) \)

IID case: \( \text{Var}(\bar{y}) = \frac{\sigma^2}{n} \) \( \hat{\text{Var}}(\bar{y}) = \frac{s^2}{n} \)

What if \( y_i \) are correlated?

\[
\text{Var}(\bar{y}) = \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right)
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(y_i, y_j)
\]

\[
= \frac{1}{n^2} \left( \sum_{i=1}^{n} \text{Var}(y_i) + \sum_{i<j} \text{Cov}(y_i, y_j) \right)
\]

\[
= \frac{1}{n^2} \left( n \sigma^2 + n(n-1) \rho \hat{\sigma}^2 \right) = \frac{1}{n} \sigma^2 \left( 1 + (n-1) \rho \right)
\]
The true variance is \( (1 + (n-1)p) \) times larger than in the iid case.

Suppose \( p = 0.1 \) (so \( R^2 \) of \( y_i \) on \( y_j \) is 0.01)

\[ N = 81 \]

\[ 1 + (n-1)p = 1 + 80 	imes 0.1 = 9 \]

True variance is 9 times the default

and the true std. error of \( \bar{y} \) is 3 times the default.
0-clusters with clustered errors

G clusters with Ng individuals per cluster

\[ y_{ig} = x_{ig}' \beta + u_{ig} \quad i = 1, \ldots, Ng, \quad g = 1, \ldots, G \]

\[ y_g = x_g' \beta + u_g \quad g = 1, \ldots, G \]

\[ y = X' \beta + u \]

\[ \hat{\beta}_{ou} = (X'X)^{-1} X' y \]

\[ = \left( \sum_{g=1}^{G} x_g x_g' \right)^{-1} \sum_{g=1}^{G} x_g y_g \]

\[ = \left( \sum_{g=1}^{G} \sum_{i=1}^{Ng} x_{ig} x_{ig}' \right)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{Ng} x_{ig} y_{ig} \]
As usual
\[ \hat{\beta} = \beta + (X'X)^{-1}X'u \quad \text{if} \quad y = X\beta + u \]

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

For simplicity, hold $X$ fixed
\[ = (X'X)^{-1} \text{Var} (X'u) (X'X)^{-1} \]
\[ = (X'X)^{-1} \text{Var} \left( \sum_{g=1}^{G} X_{g}u_{g} \right) (X'X)^{-1} \]

We assume independence of $u_g$ across $g$
\[ = (X'X)^{-1} \sum_{g=1}^{G} \text{Var}(X_{g}u_{g}) (X'X)^{-1} \]
- Suppose we have equicorrelation within cluster of errors

\[ \text{Cor}(u_{ij}, u_{jg} \mid x_i) = \begin{cases} 1 & i=j \\ \rho_n & i \neq j \end{cases} \]

This arises in a random effects model with

\[ u_{ijg} = \alpha_g + \epsilon_{ig} \]

where \( \alpha_g \) is iid and \( \epsilon_{ig} \) is iid.

- Then the incorrect OLS variance estimate \( s^2(X'X)^{-1} \) should be inflated by

\[ \tilde{\tau}_j \approx 1 + \rho_{xj} \rho_n (\bar{N}_g - 1) \]

where \( \rho_{xj} \) is within cluster correlation of \( x_j \)

\( \bar{N}_g \) is average cluster size.


CPS data example: cross-section grouped by state. 
\[ \tilde{N}_g = 81, \quad \rho_x = 1 \quad \text{and} \quad \rho_u = 0.1 \]

\[ \varepsilon_j = 1 + \rho_x \cdot \rho_u (\tilde{N}_g - 1) = 1 + 1 \times 1 \times 80 = 9. \]

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

He did not use cluster-robust SE's.

He assumed a random effects model.
Cluster Robust Variance Matrix Estimate

Outs with heteroskedastic independent errors

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

long time problem - here are N \( E \Sigma u_i^2 / x_i'x_i \) to estimate and only N observations.

White (1980) we only need a consistent estimate of the \( K \times K \) matrix

\[ \frac{1}{N} \sum_{i=1}^{N} E \Sigma u_i^2 x_i'x_i \]

Showed that we can use \[ \frac{1}{N} \left( \sum_{i=1}^{N} \hat{u}_i^2 x_i'x_i \right) \] if \( N \rightarrow \infty \).

\[ \text{Var}(\hat{\beta}) = (x'x)^{-1} \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i x_i'x_i (x'x)^{-1} \hat{u}_i = y_i - x_i\hat{\beta}_0 \]
Cluster case

\[
\text{Avar}(\hat{\beta}) = \left( \frac{1}{g} \sum_{g=1}^{6} E[x'x] \right)^{-1} \sum_{g=1}^{6} E[x'x] (E[x'x])^{-1}
\]

Old approach have a model for \( \text{Var}(u_g) = E(u_gu_g') = \Sigma_g \)

Estimate the parameters of this model giving \( \Sigma_g \rightarrow \Sigma_g \)\n
Use \( \text{Avar}(\hat{\beta}) = (x'x)^{-1} \sum_{g=1}^{6} x_g' \Sigma_g x_g (x'x)^{-1} \)

But since need to assume this model is correct

do feasible GLS

\[
\hat{\beta}_{FGLS} = \left( x' \sum x' \right)^{-1} x' \sum y, \quad \Sigma = \left[ \begin{array}{c} \Sigma_1 \\ \vdots \\ \Sigma_6 \end{array} \right]
\]

\[
\hat{\beta} = \left( \begin{array}{c} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_6 \end{array} \right)
\]

\[\text{etreg y x, re}\]
Cluster Robust estimate (CRUE)

\[ \frac{1}{G} \sum_{g=1}^{G} \tilde{\mathbf{x}}_g \tilde{\mathbf{u}}_g \tilde{\mathbf{u}}_g' \tilde{\mathbf{x}}_g - \frac{1}{G} \sum_{g=1}^{G} \mathbb{E}(\tilde{\mathbf{x}}_g' \tilde{\mathbf{u}}_g \tilde{\mathbf{u}}_g' \tilde{\mathbf{x}}_g)^p \to 0 \] as \( G \to \infty \)

So we have

\[ \text{Avar}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \frac{1}{G} \sum_{g=1}^{G} \tilde{\mathbf{x}}_g' \tilde{\mathbf{u}}_g \tilde{\mathbf{u}}_g' \tilde{\mathbf{x}}_g \left( \tilde{\mathbf{x}}_g' \tilde{\mathbf{x}}_g \right) \]

Stata uses \( \tilde{\mathbf{u}}_g = \mathbf{A} \tilde{\mathbf{u}}_g \) where \( \mathbf{A} = \frac{G \times \frac{N-1}{G-1}}{N-K} \)

where
- \( G \) \# Cluster
- \( N \) \# observations
- \( k \) \# regressors