Gov 2000: 13. Panel Data and Clustering

Matthew Blackwell

Fall 2016

1. Panel Data
2. First Differencing Methods
3. Fixed Effects Methods
4. Clustering
5. What's next for you?

Where are we? Where are we going?

- Up until now: the linear regression model, its assumptions, and violations of those assumptions
- This week: what can we do with panel data?

1/ Panel Data

Motivation

Is Democracy Good for the Poor?

Michael Ross University of California, Los Angeles

- Relationship between democracy and infant mortality?
- Compare levels of democracy with levels of infant mortality, but...
- Democratic countries are different from non-democracies in ways that we can't measure?
  - they are richer or developed earlier
  - provide benefits more efficiently
  - posses some cultural trait correlated with better health outcomes
- If we have data on countries over time, can we make any progress in spite of these problems?

Ross data

```
ross <- foreign::read.dta("../data/ross-democracy.dta")
head(ross[, c("cty_name", "year", "democracy", "infmort_unicef
```

```
##       cty_name year democracy infmort_unicef
## 1 Afghanistan 1965         0            230
## 2 Afghanistan 1966         0             NA
## 3 Afghanistan 1967         0             NA
## 4 Afghanistan 1968         0             NA
## 5 Afghanistan 1969         0             NA
## 6 Afghanistan 1970         0            215
```

Notation for panel data

- Units, $i = 1, \ldots, n$
- Time, $t = 1, \ldots, T$
- Time is a typical application, but applies to other groupings:
  - counties within states
  - states within countries
  - people within coutries, etc.
- Panel data: large n, relatively short T
- Time series, cross-sectional (TSCS) data: smaller n, large T (a political science term, mostly)

Model

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

- $\mathbf{x}_{it}$ is a vector of covariates (possibly time-varying)
- a_i is an unobserved time-constant unit effect ("fixed effect")
- u_{it} are the unobserved time-varying "idiosyncratic" errors
- $v_{it} = a_i + u_{it}$ is the combined unobserved error:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}$$

- Assume that if we could measure a_i, we would have the right model:

$$\mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$$

  - Note that this implies, u_{it} uncorrelated with $\mathbf{x}_{it}$, so that $\mathbb{E}[u_{it}|\mathbf{x}_{it}] = 0$.

Pooled OLS

- Pooled OLS: pool all observations into one regression
- Treats all unit-periods (each it) as an iid unit.
- Has two problems:
  1. Variance is wrong
  2. Possible violation of zero conditional mean errors
- Both problems arise out of ignoring the unmeasured heterogeneity inherent in a_i

Pooled OLS with Ross data

```
pooled.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur),
                 data = ross)
summary(pooled.mod)
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.7640     0.3449    28.3   <2e-16 ***
## democracy    -0.9552     0.0698   -13.7   <2e-16 ***
## log(GDPcur)  -0.2283     0.0155   -14.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.795 on 646 degrees of freedom
##   (5773 observations deleted due to missingness)
## Multiple R-squared:  0.504,  Adjusted R-squared:  0.503
## F-statistic:  329 on 2 and 646 DF,  p-value: <2e-16
```

Unmeasured heterogeneity

- If unit-effect, a_i is uncorrelated with $\mathbf{x}_{it}$, no problem for consistency!
  - $\rightsquigarrow \mathbb{E}[v_{it}|\mathbf{x}_{it}] = \mathbb{E}[a_i + u_{it}|\mathbf{x}_{it}] = 0$.
  - Just run pooled OLS (but worry about SEs).
- But a_i often correlated with $\mathbf{x}_{it}$ so that $\mathbb{E}[a_i|\mathbf{x}_{it}] \neq 0$.
  - Example: democratic institutions correlated with unmeasured aspects of health outcomes, like quality of health system or a lack of ethnic conflict.
  - Ignore the heterogeneity $\rightsquigarrow$ correlation between the combined error and the independent variables.
  - $\rightsquigarrow \mathbb{E}[v_{it}|\mathbf{x}_{it}] = \mathbb{E}[a_i + u_{it}|\mathbf{x}_{it}] \neq 0$
- Pooled OLS will be biased and inconsistent because zero conditional mean error fails for the combined error.

Panel data

- Panel data (sometimes) allows us to estimate coefficients consistently even when zero conditional mean error is violated.
- Two approaches that leverage repeated observations:
  - Differencing: look at changes over time.
  - Fixed effects: look at relationships within units.
- These approaches can help address time-constant unmeasured confounding.

2/ First Differencing Methods

First differencing

- One approach: compare changes over time
- Intuitively, changes over time will be free of time-constant unobserved heterogeneity
- Two time periods:

$$y_{i1} = \mathbf{x}'_{i1}\boldsymbol{\beta} + a_i + u_{i1}$$
$$y_{i2} = \mathbf{x}'_{i2}\boldsymbol{\beta} + a_i + u_{i2}$$

- Look at the change in y over time:

$$\begin{aligned}\Delta y_i &= y_{i2} - y_{i1} \\ &= (\mathbf{x}'_{i2}\boldsymbol{\beta} + a_i + u_{i2}) - (\mathbf{x}'_{i1}\boldsymbol{\beta} + a_i + u_{i1}) \\ &= (\mathbf{x}'_{i2} - \mathbf{x}'_{i1})\boldsymbol{\beta} + (a_i - a_i) + (u_{i2} - u_{i1}) \\ &= \Delta\mathbf{x}'_i\boldsymbol{\beta} + \Delta u_i\end{aligned}$$

First differences model

$$\Delta y_i = \Delta \mathbf{x}_i' \boldsymbol{\beta} + \Delta u_i$$

- Coefficient on the levels $\mathbf{x}_{it}$ = the coefficient on the changes $\Delta \mathbf{x}_i$
- Time-constant unobserved heterogeneity a_i drops out
- Zero conditional mean error: $\mathbb{E}[\Delta u_i | \Delta \mathbf{x}_i] = 0$ and zero conditional mean error holds.
  - Stronger than $\mathbb{E}[u_{it} | \mathbf{x}_{it}, a_i]$ because requires assumptions about relationships between u_{i2} and $\mathbf{x}_{i1}$.
- No perfect collinearity: $\mathbf{x}_{it}$ has to change over time for some units
- Under these modified assumptions, we can run regular OLS on the differences

First differences in R

```
library(plm)
fd.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross,
    index = c("id", "year"), model = "fd")
summary(fd.mod)
```

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = "fd", index = c("id", "year"))
##
## Unbalanced Panel: n=166, T=1-7, N=649
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -0.9060 -0.0956  0.0468  0.1410  0.3950
##
## Coefficients :
##             Estimate Std. Error t-value Pr(>|t|)
## (intercept)  -0.1495     0.0113  -13.26   <2e-16 ***
## democracy    -0.0449     0.0242   -1.85    0.064 .
## log(GDPcur)  -0.1718     0.0138  -12.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    23.5
## Residual Sum of Squares: 17.8
## R-Squared      :  0.246
##       Adj. R-Squared :  0.244
## F-statistic: 78.1367 on 2 and 480 DF, p-value: <2e-16
```

Differences-in-differences

- Often called "diff-in-diff", it is a special kind of FD model
- Let x_{it} be an indicator of a unit being "treated" at time t.
- Focus on two-periods where:
  - $x_{i1} = 0$ for all i
  - $x_{i2} = 1$ for the "treated group"
- Here is the basic model:

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + a_i + u_{it}$$

- d_t is a dummy variable for the second time period
  - $d_2 = 1$ and $d_1 = 0$
- β_1 is the quantity of interest: it's the effect of being treated

Diff-in-diff mechanics

- Let's take differences:

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

- $(x_{i2} - x_{i1}) = 1$ only for the treated group
- $(x_{i2} - x_{i1}) = 0$ only for the control group
- δ_0: the difference in the average outcome from period 1 to period 2 in the untreated group
- β_1 represents the additional change in y over time (on top of δ_0) associated with being in the treatment group.

Diff-in-diff interpretation

- Key idea: comparing the changes over time in the control group to the changes over time in the treated group.
- The differences between these differences is our estimate of the causal effect:

$$\beta_1 = \overline{\Delta y}_{\text{treated}} - \overline{\Delta y}_{\text{control}}$$

- Why more credible than simply looking at the treatment/control differences in period 2?

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2}$$

- a_i might be correlated with the treatment
- Unmeasured reasons why the treated group has higher or lower outcomes than the control group
- $\rightsquigarrow$ bias due to violation of zero conditional mean error

Example: Lyall (2009)

Journal of Conflict Resolution
Volume 53 Number 3
June 2009 331-362

10.1177/0022002708330881
http://jcr.sagepub.com
hosted at
http://online.sagepub.com

Does Indiscriminate Violence Incite Insurgent Attacks?

Evidence from Chechnya

Jason Lyall
Department of Politics and the Woodrow Wilson School
Princeton University, New Jersey

Example: Lyall (2009)

- Does Russian shelling of villages cause insurgent attacks?

$$\text{attacks}_{it} = \beta_0 + \delta_0 d_t + \beta_1 \text{shelling}_{it} + a_i + u_{it}$$

- We might think that artillery shelling by Russians is targeted to places where the insurgency is the strongest
- That is, part of the village fixed effect, a_i might be correlated with whether or not shelling occurs, x_{it}
- This would cause our pooled estimates to be biased
- Instead Lyall takes a diff-in-diff approach: compare attacks over time for shelled and non-shelled villages:

$$\Delta\text{attacks}_i = \delta_0 + \beta_1 \Delta\text{shelling}_i + \Delta u_i$$

Example: Card Kreuger (2009)

- Do increases to the minimum wage depress employment at fast-food restaurants?

  $$\text{employment}_{it} = \beta_0 + \delta_0 d_t + \beta_1 \text{minimum wage}_{it} + a_i + c_t + u_{it}$$

- Each i here is a different fast food restaurant in either New Jersey or Pennsylvania
- Between $t = 1$ and $t = 2$ NJ raised its minimum wage
- Employment in fast food might be driven by other state-level policies correlated with minimum wage
- Diff-in-diff approach: regress changes in employment on store being in NJ

$$\Delta \text{employment}_i = \delta_0 + \beta_1 NJ_i + \Delta u_i$$

- NJ_i indicates which stores received the treatment of a higher minimum wage at time period $t = 2$

Threats to identification

- Treatment needs to be independent of the idiosyncratic shocks:

$$\mathbb{E}[(u_{i2} - u_{i1})|(x_{i2} - x_{i1})] = \mathbb{E}[(u_{i2} - u_{i1})|x_{i2}] = 0$$

- Parallel trends: absent treatment, treated and control groups would see the same changes over time.
- Ashenfelter's dip: people who enroll in job training programs see their earnings decline prior to that training
- Lyall paper: insurgent attacks might be falling where there is shelling because rebels attacked and moved on.
- Could add covariates, sometimes called "regression diff-in-diff"

$$y_{i2} - y_{i1} = \delta_0 + \mathbf{z}_i'\tau + \beta(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

3/ Fixed Effects Methods

Fixed effects models

- Fixed effects estimation: alternative way to remove unmeasured heterogeneity
- Focuses on within-unit comparisons: changes in y_{it} and x_{it} relative to their within-group means
- First note that taking the average of the y's over time for a given unit leaves us with a very similar model:

$$\begin{aligned}
\bar{y}_i &= \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}\right] \\
&= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}'\right)\boldsymbol{\beta} + \frac{1}{T}\sum_{t=1}^{T}a_i + \frac{1}{T}\sum_{t=1}^{T}u_{it} \\
&= \bar{\mathbf{x}}_i'\boldsymbol{\beta} + a_i + \bar{u}_i
\end{aligned}$$

- Key fact: mean of the time-constant a_i is just a_i
- This regression is sometimes called the "between regression"

Within transformation

- The "fixed effects," "within," or "time-demeaning" transformation is when we subtract off the over-time means from the original data:

$$(y_{it} - \bar{y}_i) = (\mathbf{x}'_{it} - \bar{\mathbf{x}}'_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i)$$

- If we write $\ddot{y}_{it} = y_{it} - \bar{y}_i$, then we can write this more compactly as:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}'_{it}\boldsymbol{\beta} + \ddot{u}_{it}$$

Fixed effects with Ross data

```
fe.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross,
    index = c("id", "year"), model = "within")
summary(fe.mod)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = "within", index = c("id", "year"))
##
## Unbalanced Panel: n=166, T=1-7, N=649
##
## Residuals :
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -0.70500 -0.11700  0.00628  0.12200  0.75700
##
## Coefficients :
##             Estimate Std. Error t-value Pr(>|t|)
## democracy    -0.1432     0.0335   -4.28 0.000023 ***
## log(GDPcur)  -0.3752     0.0113  -33.12  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    81.7
## Residual Sum of Squares: 23
## R-Squared      :  0.718
##       Adj. R-Squared :  0.532
## F-statistic: 613.481 on 2 and 481 DF, p-value: <2e-16
```

Strict exogeneity

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{u}_{it}$$

- To use OLS on demeaned data, need $\mathbb{E}[\ddot{u}_{it}|\ddot{\mathbf{x}}_{it}] = 0$.
- This is not implied by $\mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$.
  - Only implies u_{it} will be uncorrelated with $\mathbf{x}_{it}$.
  - Need u_{it} to be uncorrelated with all $\mathbf{x}_{is}$
  - Why? $\ddot{u}_{it}$ and $\ddot{\mathbf{x}}_{it}$ are functions of errors/covariates in all time periods.
- Typical sufficient assumption is strict exogeneity:

$$\mathbb{E}[u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, a_i] = \mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$$

  - u_{it} uncorrelated with all covariates for unit i at any point in time.
  - Rules out lagged dependent variables, since $y_{i,t-1}$ has to be correlated with $u_{i,t-1}$.

Fixed effects and time-invariant covariates

- What if there is a covariate that doesn't vary over time?
  - $\rightsquigarrow x_{it} = \bar{x}_i$ and $\ddot{x}_{it} = 0$ for all periods t.
- If $\ddot{x}_{it} = 0$ for all i and t, violates no perfect collinearity.
  - R/Stata and the like will drop it from the regression.
  - Basic message: any time-constant variable gets "absorbed" by the fixed effect.
- Can include interactions between time-constant and time-varying variables, but lower order term of the time-constant variables get absorbed by fixed effects too

Time-constant variables

- Pooled model with a time-constant variable, proportion Islamic:

```
library(lmtest)
p.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
    data = ross, index = c("id", "year"), model = "pooling")
coeftest(p.mod)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30608    0.35952   28.67  < 2e-16 ***
## democracy   -0.80234    0.07767  -10.33  < 2e-16 ***
## log(GDPcur) -0.25497    0.01607  -15.87  < 2e-16 ***
## islam        0.00343    0.00091    3.77  0.00018 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time-constant variables

- FE model, where the islam variable drops out, along with the intercept:

```
fe.mod2 <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
    data = ross, index = c("id", "year"), model = "within")
coeftest(fe.mod2)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## democracy    -0.1297     0.0359   -3.62  0.00033 ***
## log(GDPcur)  -0.3800     0.0118  -32.07  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Least squares dummy variable

- Running vanilla OLS on demeaned data fine for point estimates, slightly wrong for SEs.
  - OLS doesn't know you "used" the data once to estimate the within-unit means.
- As an alternative to the within transformation, we can also include a series of $n-1$ dummy variables for each unit:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + d1_i\alpha_1 + d2_i\alpha_2 + \cdots + dn_i\alpha_n + u_{it}$$

  - Here, $d1_i$ is a binary variable which is 1 if $i = 1$ and 0 otherwise—just a unit dummy.
  - Gives the exact same point estimates as within transformation.
- Advantage: easy to implement and gives correct SEs
- Disadvantage: computationally difficult with large n, since we have to run a regression with $n + k$ variables.

Example with Ross data

```
library(lmtest)
lsdv.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur) + as.factor(id),
    data = ross)
coeftest(lsdv.mod)[1:6, ]
```

```
##                  Estimate Std. Error t value   Pr(>|t|)
## (Intercept)       13.7645    0.26597  51.751 1.008e-198
## democracy         -0.1432    0.03350  -4.276  2.299e-05
## log(GDPcur)       -0.3752    0.01133 -33.123 3.495e-126
## as.factor(id)AGO   0.2997    0.16768   1.787  7.449e-02
## as.factor(id)ALB  -1.9310    0.19014 -10.155  4.393e-22
## as.factor(id)ARE  -1.8763    0.17021 -11.024  2.387e-25
```

```
coeftest(fe.mod)[1:2, ]
```

```
##              Estimate Std. Error t value   Pr(>|t|)
## democracy    -0.1432    0.03350  -4.276  2.299e-05
## log(GDPcur)  -0.3752    0.01133 -33.123 3.495e-126
```

Fixed effects versus first differences

- Key assumptions:
  - Strict exogeneity: $E[u_{it}|\mathbf{X}_i, a_i] = 0$
  - Time-constant unmeasured heterogeneity, a_i
- Together $\implies$ fixed effects and first differences are unbiased and consistent
- With $T = 2$ the estimators produce identical estimates
- So which one is better when $T > 2$? Which one is more efficient?
  - u_{it} uncorrelated $\rightsquigarrow$ FE is more efficient
  - $u_{it} = u_{i,t-1} + e_{it}$ with e_{it} iid (random walk) $\rightsquigarrow$ FD is more efficient.
  - In between, not clear which is better.
- Large differences between FE and FD should make us worry about assumptions

4/ Clustering

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
  - Randomly assign households to different treatment conditions.
  - But the measurement of turnout is at the individual level.
- Zero conditional mean error holds here (random assignment)
- Violation of iid/random sampling:
  - errors of individuals within the same household are correlated.
  - SEs are going to be wrong.
- Called clustering or clustered dependence

Clustered dependence: notation

- Clusters (groups): $g = 1, \ldots, m$
- Units: $i = 1, \ldots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:
  - voters in households
  - individuals in states
  - students in classes
  - rulings in judges
- Outcome varies at the unit-level, y_{ig} and the main independent variable varies at the cluster level, x_g.
- Ignoring clustering is "cheating": units not independent

Clustered dependence: example model

$$
\begin{aligned}
y_{ig} &= \beta_0 + \beta_1 x_g + v_{ig} \\
&= \beta_0 + \beta_1 x_g + a_g + u_{ig}
\end{aligned}
$$

- a_g cluster error component with $\mathbb{V}[a_g|x_g] = \sigma_a^2$
- u_{ig} unit error component with $\mathbb{V}[u_{ig}|x_g] = \sigma_u^2$
- a_g and u_{ig} are assumed to be independent of each other.
  - $\rightsquigarrow \mathbb{V}[v_{ig}|x_{ig}] = \sigma_a^2 + \sigma_u^2$
- What if we ignore this structure and just use v_{ig} as the error?

Lack of independence

- Covariance between two units i and s in the same cluster:

$$\text{Cov}[v_{ig}, v_{sg}] = \sigma_a^2$$

- Correlation between units in the same group is called the intra-class correlation coefficient, or ρ_c:

$$\text{Cor}[v_{ig}, v_{sg}] = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} = \rho_c$$

- Zero covariance of two units i and s in different clusters g and k:

$$\text{Cov}[v_{ig}, v_{sk}] = 0$$

Example covariance matrix

- $\mathbf{v}' = \begin{bmatrix} v_{1,1} & v_{2,1} & v_{3,1} & v_{4,2} & v_{5,2} & v_{6,2} \end{bmatrix}$
- Variance matrix under clustering:

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \begin{bmatrix} \sigma_a^2 + \sigma_u^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 + \sigma_u^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_a^2 + \sigma_u^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 + \sigma_u^2 & \sigma_a^2 \\ 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_u^2 \end{bmatrix}$$

- Variance matrix under i.i.d.:

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \begin{bmatrix} \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_u^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_u^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_u^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_u^2 \end{bmatrix}$$

Effects of clustering

$$y_{ig} = \beta_0 + \beta_1 x_g + v_g + u_{ig}$$

- Let $\mathbb{V}_c[\widehat{\beta}_1]$ be the conventional OLS variance assuming i.i.d./homoskedasticity.
- Let $\mathbb{V}[\widehat{\beta}_1]$ be the true sampling variance under clustering.
- Relationship between the variances with equal-sized clusters clusters are balanced, $n^* = n_g$:

$$\frac{\mathbb{V}[\widehat{\beta}_1]}{\mathbb{V}_c[\widehat{\beta}_1]} \approx 1 + (n^* - 1)\,\rho_c$$

- True variance will be higher than conventional when within-cluster correlation is positive, $\rho_c > 0$.

Correcting for clustering

1. "Random effects" models (take above model as true and estimate σ_a^2 and σ_u^2)
2. Cluster-robust ("clustered") standard errors
3. Aggregate data to the cluster-level and use OLS
   $\bar{y}_g = \frac{1}{n_g} \sum_i y_{ig}$
   - If n_g varies by cluster, then cluster-level errors will have heteroskedasticity
   - Can use WLS with cluster size as the weights

Cluster-robust SEs

- First, let's write the within-cluster regressions like so:

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{v}_g$$

- $\mathbf{y}_g$ is the vector of responses for cluster g, and so on
- We assume that respondents are independent across clusters, but possibly dependent within clusters. Thus, we have

$$\mathbb{V}[\mathbf{v}_g | \mathbf{X}_g] = \Sigma_g$$

- Remember our sandwich expression:

$$\mathbb{V}[\widehat{\boldsymbol{\beta}} | \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \Sigma \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

- Under this clustered dependence, we can write this as:

$$\mathbb{V}[\widehat{\boldsymbol{\beta}} | \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^{m} \mathbf{X}_g' \Sigma_g \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

Estimating CRSEs

- Way to estimate this matrix: replace Σ_g with an estimate based on the within-cluster residuals, $\hat{\mathbf{v}}_g$:

$$\widehat{\Sigma}_g = \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g'$$

- Final expression for our cluster-robust covariance matrix estimate:

$$\widehat{\mathbb{V}}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^{m} \mathbf{X}_g' \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

- With small-sample adjustment (which is what most software packages report):

$$\widehat{\mathbb{V}}_a[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{m}{m-1} \frac{n-1}{n-k-1} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^{m} \mathbf{X}_g' \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

Example: Gerber, Green, Larimer

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

Social pressure model

```
load("../data/gerber_green_larimer.RData")
library(lmtest)
social$voted <- 1 * (social$voted == "Yes")
social$treatment <- factor(social$treatment, levels = c("Control",
    "Hawthorne", "Civic Duty", "Neighbors", "Self"))
mod1 <- lm(voted ~ treatment, data = social)
coeftest(mod1)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.29664    0.00106  279.53  < 2e-16 ***
## treatmentHawthorne  0.02574    0.00260    9.90  < 2e-16 ***
## treatmentCivic Duty 0.01790    0.00260    6.88  5.8e-12 ***
## treatmentNeighbors  0.08131    0.00260   31.26  < 2e-16 ***
## treatmentSelf       0.04851    0.00260   18.66  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Social pressure model, CRSEs

- No canned CRSE in R, we posted some code on Canvas:

```
source("vcovCluster.R")
coeftest(mod1, vcov = vcovCluster(mod1, "hh_id"))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.29664    0.00131  226.52  < 2e-16 ***
## treatmentHawthorne    0.02574    0.00326    7.90  2.8e-15 ***
## treatmentCivic Duty   0.01790    0.00324    5.53  3.2e-08 ***
## treatmentNeighbors    0.08131    0.00337   24.13  < 2e-16 ***
## treatmentSelf         0.04851    0.00330   14.70  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cluster-robust standard errors

- CRSE do not change our estimates $\widehat{\boldsymbol{\beta}}$, cannot fix bias
- CRSE is consistent estimator of $\mathbb{V}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ given clustered dependence
  - Relies on independence between clusters
  - Allows for arbitrary dependence within clusters
  - CRSEs usually > conventional SEs—use when you suspect clustering
- Consistency of the CRSE are in the number of groups, not the number of individuals
  - CRSEs can be incorrect with a small (< 50 maybe) number of clusters

5/ What's next for you?

Where are you?

- You've been given a powerful set of tools

Your new weapons

- Probability: if we knew the true parameters (means, variances, coefficients), what kind of data would we see?
- Inference: what can we learn about the truth from the data we have?
- Regression: how can we learn about relationships between variables?

You need more training!

- We got through a ton of solid foundation material, but to be honest, we have basically got you to the state of the art in political science in the 1970s

What else to learn?

- Non-linear models (Gov 2001)
  - what if y_i is not continuous?
- Maximum likelihood (Gov 2001)
  - a general way to do inference and derive estimators for almost any model
- Bayesian statistics (Stat 120/220)
  - an alternative approach to inference based on treating parameters as random variables
- Causal inference (Gov 2002, Stat 186)
  - how do we make more plausible causal inferences?
  - what happens when treatment effects are not constant?

Glutton for punishment?

- Stat 110/111: rigorous introduction to probability and inference
- Stat 210/211: Stats PhD level introduction to probability and inference (measure theory)
- Stat 221: statistical computing

Thanks!

Fill out your evaluations!