

# 14. Panel and Clustered Data

Spring 2023

Matthew Blackwell

Gov 2002 (Harvard)

# Where are we? Where are we going?

- Focus up until now on iid data, but often doesn't hold.
- **Panel** and **clustered** data are two common non-iid data.
- Panel data also holds hope for removing unmeasured heterogeneity.

# 1/ Panel Data

## 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
  - possess 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

```
library(tidyverse)
library(haven)
ross <- read_dta("../assets/ross-democracy.dta")
ross <- ross |>
  filter(!is.na(kidmort_unicef), !is.na(democracy), !is.na(GDPcur)) |>
  group_by(id) |>
  filter(var(democracy, na.rm = TRUE) > 0)
head(ross[,c("cty_name", "year", "democracy", "infmtort_unicef")])
```

```
## # A tibble: 6 x 4
##   cty_name   year democracy infmort_unicef
##   <chr>     <dbl>     <dbl>         <dbl>
## 1 Albania   1990         0             36
## 2 Albania   1995         1             30
## 3 Argentina 1970         0             59
## 4 Argentina 1980         0             33
## 5 Argentina 1990         1             25
## 6 Argentina 1995         1             22
```

# Notation for panel data

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

$$Y_{it} = \mathbf{X}'_{it}\boldsymbol{\beta} + c_i + u_{it}$$

- $\mathbf{X}_{it}$  is a vector of covariates (possibly time-varying)
- $c_i$  is an **unobserved** time-constant unit effect (“fixed effect”)
  - Confusingly, we’ll allow them to be random variables.
- $u_{it}$  are the unobserved time-varying “idiosyncratic” errors
- $v_{it} = c_i + u_{it}$  is the combined unobserved error:  $Y_{it} = \mathbf{X}'_{it}\boldsymbol{\beta} + v_{it}$
- Assume that if we could measure  $c_i$ , we would have the correct CEF:

$$\mathbb{E}[u_{it} \mid \mathbf{X}_{it}, c_i] = 0 \quad \implies \quad \mathbb{E}[Y_{it} \mid \mathbf{X}_{it}, c_i] = \mathbf{X}'_{it}\boldsymbol{\beta} + c_i$$

# 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 probably wrong if there is dependence over time
  2. Errors might be correlated with the covariates
- Both problems arise out of ignoring the **unmeasured heterogeneity** inherent in  $c_i$

# Pooled OLS with Ross data

```
library(lmtest)
library(sandwich)
pooled.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur),
                 data = ross)
coeftest(pooled.mod, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.3338     0.6279   16.46 < 2e-16 ***
## democracy    -0.5639     0.1135   -4.97  1.3e-06 ***
## log(GDPcur)  -0.2486     0.0287   -8.66  7.7e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Unmeasured heterogeneity

- Since  $u_{it}$  is the CEF error,  $\mathbf{X}_{it}$  are uncorrelated with it:  $\mathbb{E}[\mathbf{X}_{it}u_{it}] = 0$ .
- If unit-effect  $c_i$  is uncorrelated with  $\mathbf{X}_{it}$ , no problem for consistency!
  - $\rightsquigarrow \mathbb{E}[\mathbf{X}_{it}v_{it}] = \mathbb{E}[\mathbf{X}_{it}(c_i + u_{it})] = 0$ .
  - Just run pooled OLS (but worry about SEs).
- But  $c_i$  often correlated with  $\mathbf{X}_{it}$  so that  $\mathbb{E}[\mathbf{X}_{it}c_i] \neq 0$ .
  - 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.
  - $\rightsquigarrow \mathbb{E}[\mathbf{X}_{it}v_{it}] = \mathbb{E}[\mathbf{X}_{it}(c_i + u_{it})] = \mathbb{E}[\mathbf{X}_{it}c_i] \neq 0$
- Pooled OLS will be inconsistent for the CEF parameters,  $\beta$ .

# Strict exogeneity

- Panel data allows us to estimate  $\beta$  even in this setting
- Two approaches that leverage repeated observations:
  - **Differencing** look at changes over time.
  - **Fixed effects** look at relationships within units.

## **2/** First Differencing Methods

# First differencing

- One approach: compare **changes over time**
- Intuitively, time-constant heterogeneity can't affect changes over time.
- Two time periods:

$$Y_{i1} = \mathbf{X}'_{i1}\boldsymbol{\beta} + c_i + u_{i1}$$

$$Y_{i2} = \mathbf{X}'_{i2}\boldsymbol{\beta} + c_i + u_{i2}$$

- Look at the change in  $Y$  over time:

$$\begin{aligned}\Delta Y_i &= Y_{i2} - Y_{i1} \\ &= (\mathbf{X}'_{i2}\boldsymbol{\beta} + c_i + u_{i2}) - (\mathbf{X}'_{i1}\boldsymbol{\beta} + c_i + u_{i1}) \\ &= (\mathbf{X}'_{i2} - \mathbf{X}'_{i1})\boldsymbol{\beta} + (c_i - c_i) + (u_{i2} - u_{i1}) \\ &= \Delta\mathbf{X}'_i\boldsymbol{\beta} + \Delta u_i\end{aligned}$$

# 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  $c_i$  drops out.
- For consistency of OLS on the differences, we need  $\mathbb{E}[\Delta \mathbf{X}_i \Delta u_i] = 0$ .

$$\mathbb{E}[(\mathbf{X}_{i2} - \mathbf{X}_{i1})(u_{i2} - u_{i1})] = \mathbb{E}[\mathbf{X}_2 u_2] + \mathbb{E}[\mathbf{X}_1 u_1] - \mathbb{E}[\mathbf{X}_1 u_2] - \mathbb{E}[\mathbf{X}_2 u_1] = 0$$

- First two are 0 since we assume the CEF is correctly specified up to  $c_i$
- $\mathbb{E}[\mathbf{X}_1 u_2]$  and  $\mathbb{E}[\mathbf{X}_2 u_1]$  are additional assumptions: no **feedback between outcome and covariates**
- Invertibility of  $\mathbb{E}[\Delta \mathbf{X}_{it} \Delta \mathbf{X}'_{it}]$  requires  $\mathbf{X}_{it}$  to vary over time for someone
- Under these assumptions, pooled OLS on the differences is consistent.

## **3/** Fixed Effects Methods

# Fixed effects models

- **Fixed effects model:** another 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:

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

- Key fact: mean of the time-constant  $c_i$  is just  $c_i$
- This regression is sometimes called the “between regression”

# Within transformation

- **Fixed effect** or **within transformation**:

$$(Y_{it} - \bar{Y}_i) = (\mathbf{X}'_{it} - \bar{\mathbf{X}}'_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i)$$

- Center every covariate and the outcome at its within-unit mean.
  - $c_i$  drops out because its within-unit mean is itself (time-constant).
- 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

```
library(fixest)
fe.mod <- fixest::feols(
  log(kidmort_unicef) ~ democracy + log(GDPcur) | id,
  data = ross, vcov = "hetero")
summary(fe.mod)
```

```
## OLS estimation, Dep. Var.: log(kidmort_unicef)
## Observations: 237
## Fixed-effects: id: 53
## Standard-errors: Heteroskedasticity-robust
##           Estimate Std. Error t value    Pr(>|t|)
## democracy    -0.156    0.0314   -4.97 0.0000015379 ***
## log(GDPcur)   -0.354    0.0252  -14.03 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.18124      Adj. R2: 0.95396
##           Within R2: 0.711842
```

# Strict exogeneity

$$\ddot{Y}_{it} = \ddot{\mathbf{X}}_{it}'\boldsymbol{\beta} + \ddot{u}_{it}$$

- To use OLS on demeaned data, need  $\mathbb{E}[\ddot{\mathbf{X}}_{it}\ddot{u}_{it}] = 0$ .
- This is not implied by  $\mathbb{E}[u_{it}|\mathbf{X}_{it}, c_i] = 0$ .
  - Only implies  $u_{it}$  will be uncorrelated with  $\mathbf{X}_{it}$ .
  - Like with differencing, 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**.
- Key assumption is **strict exogeneity**:

$$\mathbb{E}[u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \dots, \mathbf{X}_{iT}, c_i] = 0$$

- $u_{it}$  uncorrelated with all covariates for unit  $i$  at any point in time.
- Rules out lagged dependent variables, since  $Y_{i,t-1}$  is a function of  $u_{i,t-1}$ .

# 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 invertibility.
  - R/Stata and the like will drop it from the regression.
  - 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 <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam, data = ross)
coeftest(p.mod, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.36014    0.58133   17.82 < 2e-16 ***
## democracy   -0.47634    0.09441   -5.05 9.6e-07 ***
## log(GDPcur) -0.25597    0.02671   -9.58 < 2e-16 ***
## islam        0.00855    0.00106    8.06 5.2e-14 ***
## ---
## 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 <- feols(  
  log(kidmort_unicef) ~ democracy + log(GDPcur) + islam | id,  
  data = ross, vcov = "hetero")  
summary(fe.mod2)
```

```
## OLS estimation, Dep. Var.: log(kidmort_unicef)  
## Observations: 220  
## Fixed-effects: id: 45  
## Standard-errors: Heteroskedasticity-robust  
##           Estimate Std. Error t value   Pr(>|t|)  
## democracy    -0.144     0.0347   -4.14 0.000054978  
## log(GDPcur)  -0.360     0.0257  -14.00 < 2.2e-16  
##  
## democracy    ***  
## log(GDPcur)  ***  
## ... 1 variable was removed because of collinearity (islam)  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
## RMSE: 0.185449      Adj. R2: 0.949078  
##                   Within R2: 0.717818
```

# Least squares dummy variable

- Naive OLS on demeaned data is ok for  $\hat{\beta}$  but the SEs are wrong.
  - OLS doesn't know you "used" the data to estimate the within-unit means.
- As an alternative, **dummy variable estimator** regressing:

$$Y_{it} \text{ on } \mathbf{X}_{it}, D_{i2}, D_{i3}, \dots, D_{in}$$

- Here,  $D_{i2}$  is a binary variable which is 1 if  $i = 2$  and 0 otherwise.
  - Gives the **exact** same point estimates as within transformation.
- Comments:
  - Pros: easy to implement and gives correct SEs.
  - Con: computationally slow with large  $n$ .
  - Usually better to use dedicated software like `fixest` package in R.

# Example with Ross data

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

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	11.385	0.6306	18.05	2.01e-42
## democracy	-0.156	0.0366	-4.27	3.14e-05
## log(GDPcur)	-0.354	0.0295	-11.99	8.65e-25
## idARG	1.263	0.1425	8.87	6.82e-16
## idARM	0.462	0.1287	3.59	4.20e-04
## idBEN	1.334	0.0884	15.08	7.10e-34

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

##	Estimate	Std. Error	z value	Pr(> z )
## democracy	-0.156	0.0314	-4.97	6.69e-07
## log(GDPcur)	-0.354	0.0252	-14.03	1.08e-44

## 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):  $g = 1, \dots, m$
- Units:  $i = 1, \dots, n_g$
- $n_g$  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,  $X_g$ .

# Clustered dependence: example model

$$\begin{aligned} Y_{ig} &= \beta_0 + X_g \beta_1 + v_{ig} \\ &= \beta_0 + X_g \beta_1 + c_g + u_{ig} \end{aligned}$$

- $u_{ig}$  unit error component with  $\mathbb{V}[u_{ig}|X_g] = \sigma_u^2$
- $c_g$  cluster error component with  $\mathbb{V}[c_g|X_g] = \sigma_c^2$
- $c_g$  and  $u_{ig}$  are assumed to be independent of each other.
  - $\rightsquigarrow \mathbb{V}[v_{ig}|X_g] = \sigma_c^2 + \sigma_u^2$
- 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:

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

- Correlation between units in the same group is called the **intra-class correlation coefficient**, or  $\rho_c$ :

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

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

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

# Example covariance matrix

- $\mathbf{v}' = [ v_{1,1} \quad v_{2,1} \quad v_{3,1} \quad v_{4,2} \quad v_{5,2} \quad v_{6,2} ]$
- Variance matrix under clustering:

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

- 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

$$Y_{ig} = \beta_0 + X_g \beta_1 + c_g + u_{ig}$$

- $\mathbb{V}^0[\hat{\beta}_1] =$  **conventional** OLS variance assuming i.i.d./homoskedasticity.
- Let  $\mathbb{V}[\hat{\beta}_1]$  be the true sampling variance under clustering.
- When clusters are balanced,  $n^* = n_g$ , comparison of clustered to conventional:

$$\mathbb{V}[\hat{\beta}_1] \approx \mathbb{V}^0[\hat{\beta}_1] (1 + (n^* - 1)\rho_c)$$

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

# Linear model with clustering

$$Y_{ig} = \mathbf{X}'_{ig}\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} | \mathbf{X}_{ig}] = 0$  so we have the correct CEF.
  - $\mathbb{E}[v_{ig}v_{jg'} | \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$  unless  $g = g'$ .
  - Correlated errors allowed within groups, uncorrelated across. Allows heteroskedasticity.
- Pooled OLS under clustered dependence:

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

- $\mathbf{Y}_g$  is the  $n_g \times 1$  vector of responses for cluster  $g$
  - $\mathbb{X}_g$  is the  $n_g \times k$  matrix of data for the  $g$ th cluster.
- We can write the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \left( \sum_{g=1}^m \mathbb{X}'_g \mathbb{X}_g \right) \left( \sum_{g=1}^m \mathbb{X}'_g \mathbf{Y}_g \right)$$

# Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as  $m$  gets big.
  - Key intuition: we're sampling clusters, not individual units.
- CLT implies  $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  will be asymp. normal with mean 0 and variance:

$$(\mathbb{E}[\mathbb{X}'_g \mathbb{X}_g])^{-1} \mathbb{E}[\mathbb{X}'_g \mathbf{v}_g \mathbf{v}'_g \mathbb{X}_g] (\mathbb{E}[\mathbb{X}'_g \mathbb{X}_g])^{-1}$$

- Similar to the iid case, replace population quantities with sample versions (and divide by  $m$ ):

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{CLO}} = (\mathbb{X}'\mathbb{X})^{-1} \left( \sum_{g=1}^m \mathbb{X}'_g \hat{\mathbf{v}}_g \hat{\mathbf{v}}'_g \mathbb{X}_g \right) (\mathbb{X}'\mathbb{X})^{-1}$$

- Noting:  $\mathbb{X}'\mathbb{X}/m = m^{-1} \sum_{g=1}^m \mathbb{X}'_g \mathbb{X}_g$
- With small-sample adjustment (reported by most software):

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{CL1}} = \frac{m}{m-1} \frac{n-1}{n-k} (\mathbb{X}'\mathbb{X})^{-1} \left( \sum_{g=1}^m \mathbb{X}'_g \hat{\mathbf{v}}_g \hat{\mathbf{v}}'_g \mathbb{X}_g \right) (\mathbb{X}'\mathbb{X})^{-1}$$

# 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("../assets/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
## (Intercept)    0.29664    0.00106  279.53
## treatmentHawthorne  0.02574    0.00260    9.90
## treatmentCivic Duty 0.01790    0.00260    6.88
## treatmentNeighbors  0.08131    0.00260   31.26
## treatmentSelf      0.04851    0.00260   18.66
##
##              Pr(>|t|)
## (Intercept)    < 2e-16 ***
## treatmentHawthorne < 2e-16 ***
## treatmentCivic Duty 5.8e-12 ***
## treatmentNeighbors < 2e-16 ***
## treatmentSelf    < 2e-16 ***
## ---
```

# Social pressure model, CRSEs

```
library(sandwich)
coeftest(mod1, vcov = sandwich::vcovCL(mod1, cluster = social$hh_id))
```

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

# Cluster-robust standard errors

- CRSE do not change our estimates  $\hat{\beta}$ , cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
  - Relies on independence between clusters
  - Allows for arbitrary dependence within clusters
  - CRSEs usually  $>$  conventional SEs—use when you suspect clustering
- When  $X_{ig}$  not constant within cluster, but just correlated  $\rightsquigarrow$  more complicated.
  - See Abadie, Athey, Imbens, and Wooldridge (2021).
- 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