UNIVERSITY OF ILLINOIS

AT URBANA-CHAMPAIGN

Using Panel (aka Longitudinal) Data Estimators to Identify Causal Effects

1867

illinois.edu

Outline for the Session

1. The Omitted Variables Problem (OVP)
2. Different panel estimators
3. Attrition and unbalanced panels
4. The art of the possible…
5. SUTVA Violations: Spillovers and Network Effects

The Omitted Variables Problem

The Omitted Variable Problem (OVP)

- Causal inference is a missing variables or omitted variables problem
  - We don't know what happened to those treated in the absence of the treatment
- RCTs solves the OVP by ensuring treatment and control groups are equivalent through randomization
  - We then assume the control group is representative of what would have happened to the treatment group had they not been treated

The Omitted Variable Problem (OVP)

- Matching solves the OVP by constructing a control group based on observable characteristics
  - Conditional on observables the matched group is representative of what would have happened to the treatment group had they not been treated
  - But this does not control for unobservables

The Omitted Variable Problem (OVP)

- IVs solve the OVP by assuming that there are unobservable differences between treatment and control and finding an instrument to break the correlation between the treatment and the unobservable differences
  - Conditional on a set of Identifying Assumptions the IV allows us to control for unobserved characteristics that make the treatment and control groups different and affect the outcome

The Omitted Variable Problem (OVP)

- Panel data techniques provide an additional way to try and establish causal inference
  - When we have multiple observations of plots/households/firms over time we can control for time invariant unobservables and common shocks to obtain consistent and unbiased estimates of the treatment effect

Some Preliminary Assumptions

- Assume a large population of cross-sectional units (plot, household, firm) that we can observe over time
- We randomly sample from the cross-section, so observations are necessarily independent in the cross-section
- We have a large cross-section (n) and relatively few time periods (t)

Some Preliminary Assumptions

- An individual-specific time-invariant unobservable, c_i, is drawn along with the observed data
  - *E.g. unobserved characteristics that affect probability of adoption, or for yield to be always better for one farmer than another.*
- Common shock, τ_t
  - *Prices, el nino.*

$$Y_{it} = \alpha X_{it} + \beta T_{it} + c_i + \tau_t + \epsilon_{it}$$

Some Preliminary Assumptions

$$Y_{it} = \alpha X_{it} + \beta T_{it} + c_i + \tau_t + \epsilon_{it}$$

- X_{it} is a set of observed variables that may combine variables that vary only over time (market price), over individual (soils) or both (weather).
- ϵ_{it} are the idiosyncratic errors
  - The composite error term is $v_{it} = c_i + \tau_t + \epsilon_{it}$
  - v_{it} is almost certainly serially correlated and definitely is if ϵ_{it} is serially uncorrelated. This will be because the value of c_i is the same for all t

Discussion: Irrigation Project Example

$$\log(income) = \alpha + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \tau_t + \epsilon_{it}$$

- $Irrig_{it}$ is the treatment, if the households had received the irrigation project
- $dist_i$ is household distance to market and does not change over time

Discussion: Irrigation Project Example

$$\log(income) \\ = \alpha + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \tau_t + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

Discussion: Irrigation Project Example

$$\log(income) = \alpha + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \tau_t + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

Discussion: Irrigation Project Example

$$\log(income) = \alpha + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \tau_t + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

YES

Discussion: Irrigation Project Example

$$\log(income) = \alpha + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \tau_t + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

    YES

  - Are those factors, in c_i, correlated with $Irrig_{it}$?

Discussion: Irrigation Project Example

$$\log(income) = \alpha + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \tau_t + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

YES

  - Are those factors, in c_i, correlated with $Irrig_{it}$?

Probably

Different Panel Data Models

Panel Data Models

- Primary focus will be on the following
  - Pooled Ordinary Least Squares (OLS)
  - Random Effects (RE)
  - Fixed Effects (FE)
  - Correlated Random Effects (CRE)
- Alternative models
  - First Differencing (FD)
  - Multilevel Model (MLM)

Pooled OLS

- Assumes $Cov(v_{it}, v_{is}) = 0$ and $Cov(v_{it}, v_{jt}) = 0$
  - In words: the composite error term is uncorrelated across time (no serial correlation).
  - And across individuals (and treatment groups)
  - This will clearly not be true if there are time-invariant unobserved effects in our model or group effects
- How likely is it that there are no unobserved effects in our model?
  - Back to the U's

Pooled OLS

- Using OLS estimate

$$Y_{it} = \alpha + \delta R_i + \gamma X_{it} + c_i + \tau_t + \epsilon_{it}$$

- To test if the errors are serially uncorrelated, save $\hat{\epsilon}_{it}$ and then regress
  - $\hat{\epsilon}_{it} = \rho\hat{\epsilon}_{it-1} + u_t$
  - If $\rho = 0$ then errors are serially uncorrelated and Pooled OLS is BLUE
  - If $\rho \neq 0$ then errors are serially correlated and you need a panel data estimator

Random Effects

- Assumes $Cov(X_{it}, c_i) = 0$
  - Alternatively, $E[c_i|X_{it}] = E[c_i]$ – conditional mean independence
  - In words: the unobserved effect is uncorrelated with the observed explanatory variables
- How likely is it that unobserved individual characteristics are uncorrelated with observed characteristics?
  - Isn't the whole point of using panel data to allow for c_i to be arbitrarily correlated with X_{it}?

Random Effects

- Using GLS estimate

$$Y_{it} = \alpha + \delta R_i + \gamma X_{it} + v_{it}$$

- Several tests for validity of REs
  - To test if $c_i = 0$ you can use the Breusch-Pagan Lagrangian multiplier test for RE
  - To test if the unobserved effect is uncorrelated with the observed explanatory variables we can use a Hausman Test

Fixed Effects

- Allows for $Cov(X_{it}, c_i) \neq 0$
  - Alternatively, $E[c_i|X_{it}]$ is allowed to be any value
  - In words: allows for arbitrary correlation between unobserved effect and the observed explanatory variables
  - Explicitly estimate c_i and/or τ_i
- Equivalent to 'de-meaning' the data in a linear model
- But, panel FE does not allow us to simultaneously estimate time-constant variables
  - Can back them out in a secondary regression:

$$\hat{c}_i = \alpha + \gamma X_i + \mu_i$$

Fixed Effects

- estimate

$$Y_{it} = \gamma X_{it} + \zeta c_i + \theta \tau_t + \epsilon_{it}$$

- Include binary indicators for each individual
  - Note this controls for c_i but removes R_i due to perfect collinearity

Correlated Random Effects

- Assumes $E[c_i|X_{it}] = E[c_i|\bar{X}_i] = \psi + \xi\bar{X}_i$
  - In words: we model the dependence between unobserved effect and the observed explanatory variables as

$$c_i = \psi + \xi\bar{X}_i + a_i$$

- Allows us to unify FE and RE estimation approaches

Correlated Random Effects

- First, define the relationship between the unobserved effect and the observed covariates

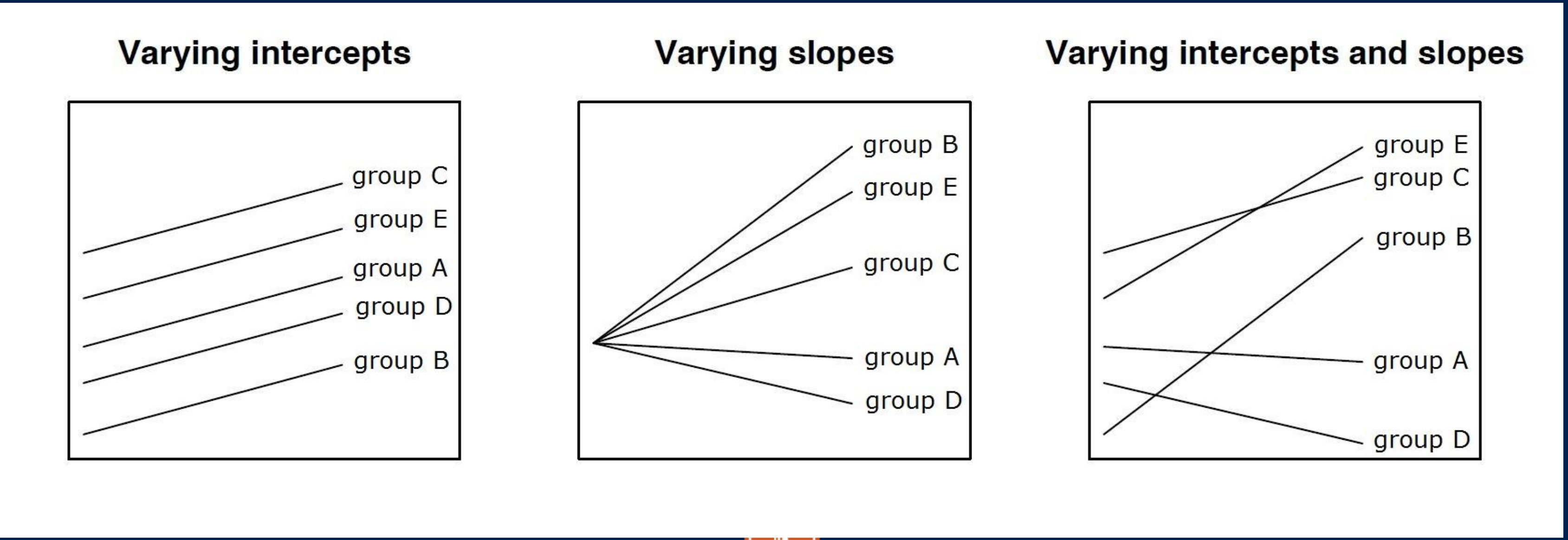
$$c_i = \psi + \xi \bar{X}_i + a_i$$

- Second, estimate the equation with OLS

$$Y_{it} = \theta G_t + \delta R_i + \gamma X_{it} + \psi + \xi \bar{X}_i + a_i + \epsilon_{it}$$

How many FE should we include?

- Individual or Group FE
- Time
- Group x time



Where is our variation coming from?

Year	a	b	c	d	Ave
1	100	120	110	140	117.5
2	110	135	105	155	125
3	85	90	100	110	96.25
4	150	140	95	145	133.75
Ave	110	122.5	102.5	137.5	

OLS – variation between households and over time

With year FE

Year	a	b	c	d	Ave
1	100 [-17.5]	120 [2.5]	110 [-7.5]	140 [22.5]	117.5
2	110 [-15]	135 [-10]	105 [-20]	155 [25]	125
3	85 [-16.25]	90 [-6.25]	100 [3.75]	110 [13.75]	96.25
4	150 [16.25]	140 [6.25]	95 [-38.75]	145 [12.25]	133.75
Ave	110	122.5	102.5	137.5	

Difference among households within year

Common shocks (e.g. world price; el nino)

With HH FE?

Year	a	b	c	d	Ave
1	100 [-10]	120 [-2.5]	110 [-7.5]	140 [2.5]	117.5
2	110 [0]	135 [22.5]	105 [2.5]	155 [17.5]	125
3	85 [-25]	90 [-32.5]	100 [2.5]	110 [-27.5]	96.25
4	150 [40]	140 [27.5]	95 [-7.5]	145 [7.5]	133.75
Ave	110	122.5	102.5	137.5	

Household-specific effects (soil type, education, farm size)

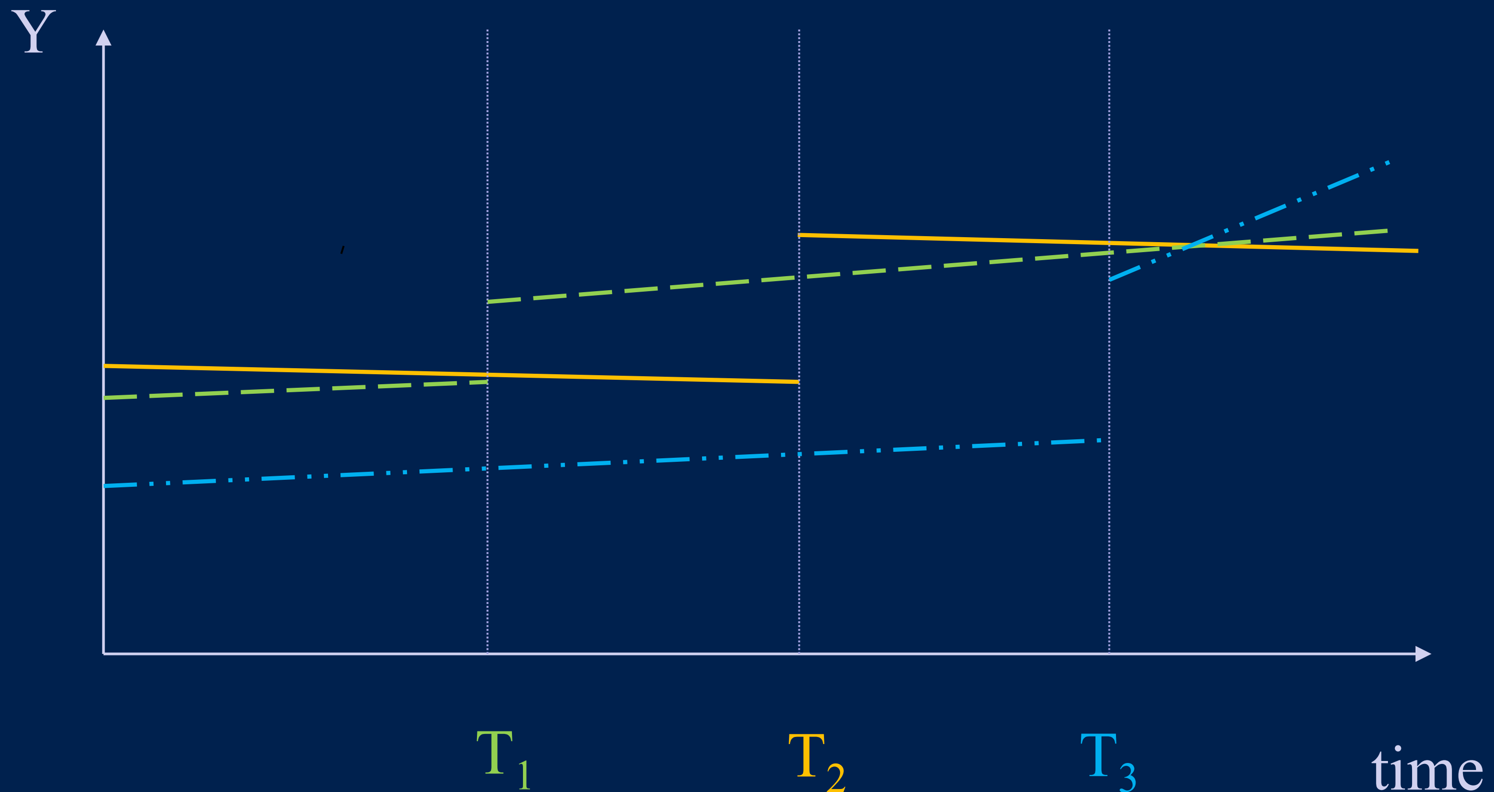
Now comparing households to themselves over time

With group time trends?

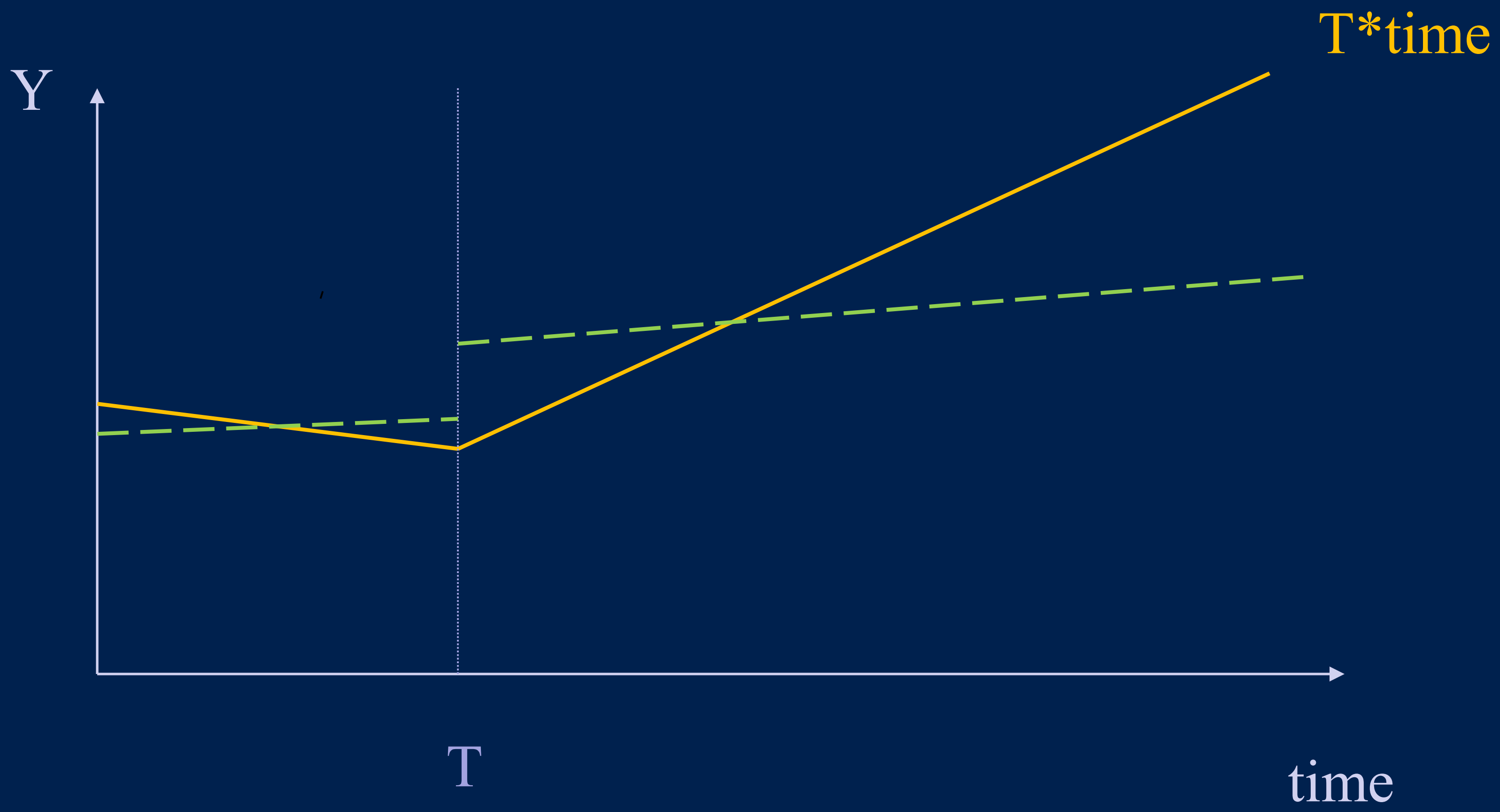
Year	a	b	c	d	Ave
1	100 [-10]	120 [-2.5]	110 [-7.5]	140 [2.5]	117.5
2	110 [0]	135 [22.5]	105 [2.5]	155 [17.5]	125
3	85 [-25]	90 [-32.5]	100 [2.5]	110 [-27.5]	96.25
4	150 [40]	140 [27.5]	95 [-7.5]	145 [7.5]	133.75
Ave	110	122.5	102.5	137.5	

Now comparing household deviation from group trend

Treatment over time



What if treatment affects trajectory, not level?



Attrition

Practical issue when collecting longitudinal (panel) data.

- Some households will be away, some will have a different respondent
- Some households will have migrated
- Some will no longer want to participate

Check %, check whether missing observations are systematically different from folks staying

Collect data on new households to preserve geographic sample

-> *Unbalanced Panel Methods*

The Art of the Possible…

So you don't have baseline data…

- Recall?
- Secondary data? (national surveys; satellite imagery)

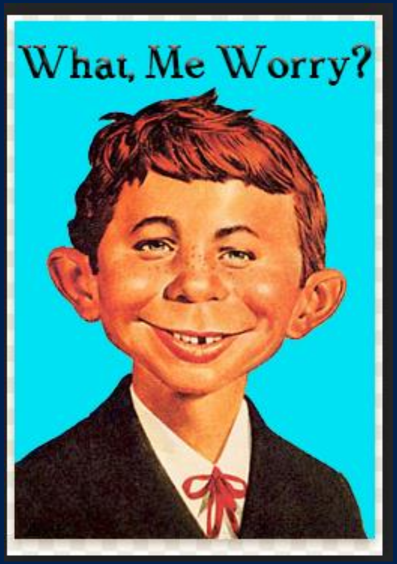
So you don't have data on controls…

- Variation in treatment intensity?
- Variation in treatment timing?

In general

- Placebo tests – rule out other options (informed by theory of change)
- Qualitative data to rule out other options

SUTVA Violations



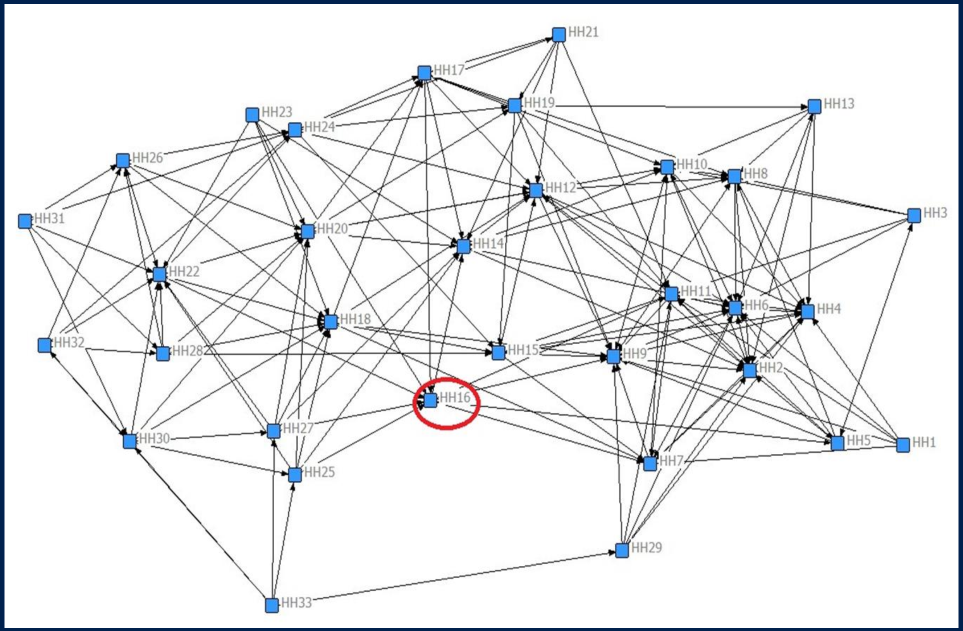
Spillovers (when SUTVA falls apart…)

- Social Networks
- Peer Effects
- Group threshold Effects
- Spatial Spillovers

- *Bias estimated treatment effects*
- *Often important in and of themselves*
- *Ideally integrate into research design*

Social Network Effects

- Where a program is placed within a social network matters
- Banerjee et al (2011) – microfinance in India
- Songersemsawas et al (2015) – contract choice



Peer Effects

- Reflection Problem
- Can solve through using characteristics of friends of friends as instruments
- Do peer effects through social networks affect…
  - Input use in new crops (Conley and Udry 2010)
  - Land allocation to new crops (Munshi 2004)
  - Market mechanisms (Fafchamps and Minton 1998, 1999, 2002; Michelson 2015)
  - Agricultural revenue (Songsermsawas et al 2015b)

Mechanism?

- Influence versus Information (Montgomery and Casterline)
- Oster and Thornton (2012)
  - *Wanting to do like friends?*
  - *Switching behavior because of friends' positive benefits?*
  - *Learning how to use a new technology*

Within village spillovers and threshold effects

Within Village Spillovers

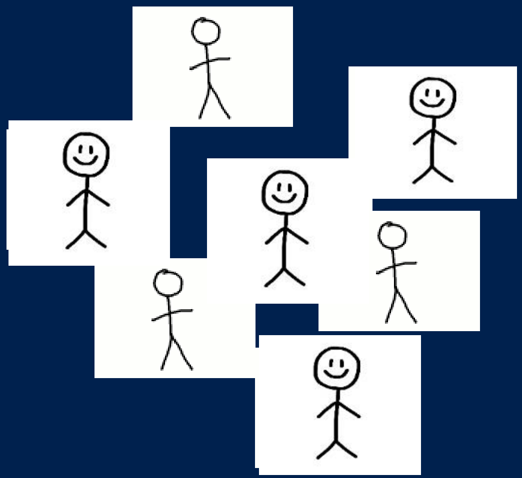
- Can identify through different intensity of treatment (Baird et al 2015)
- Can identify through modeling peer networks

Threshold Effects

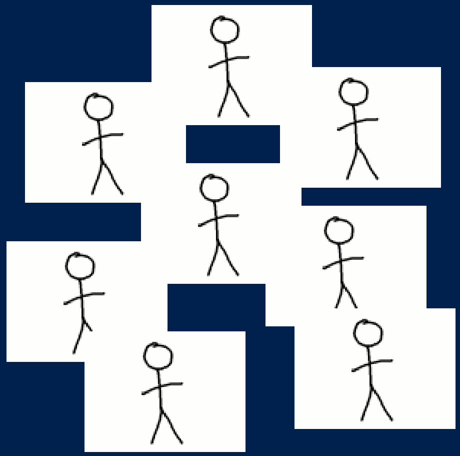
- Idea that an intervention needs to reach a certain saturation point to have an effect

example

- Only some people eligible



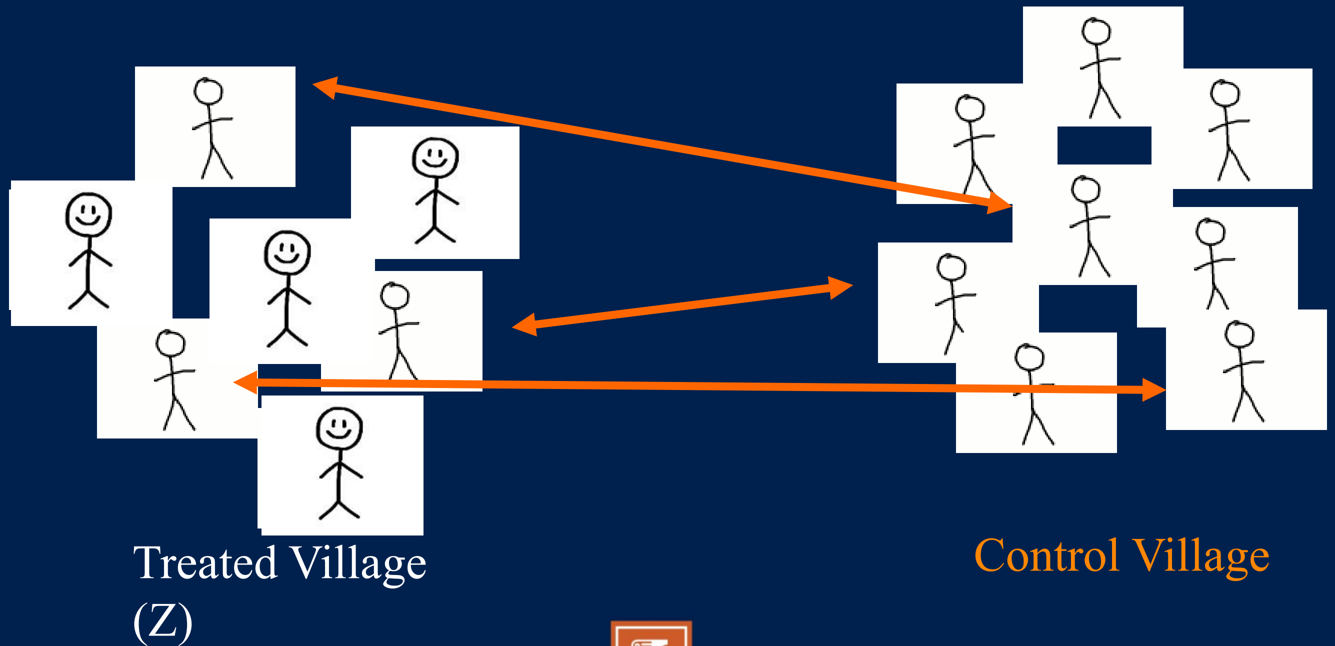
Treated Village
(Z)



Control Village

example

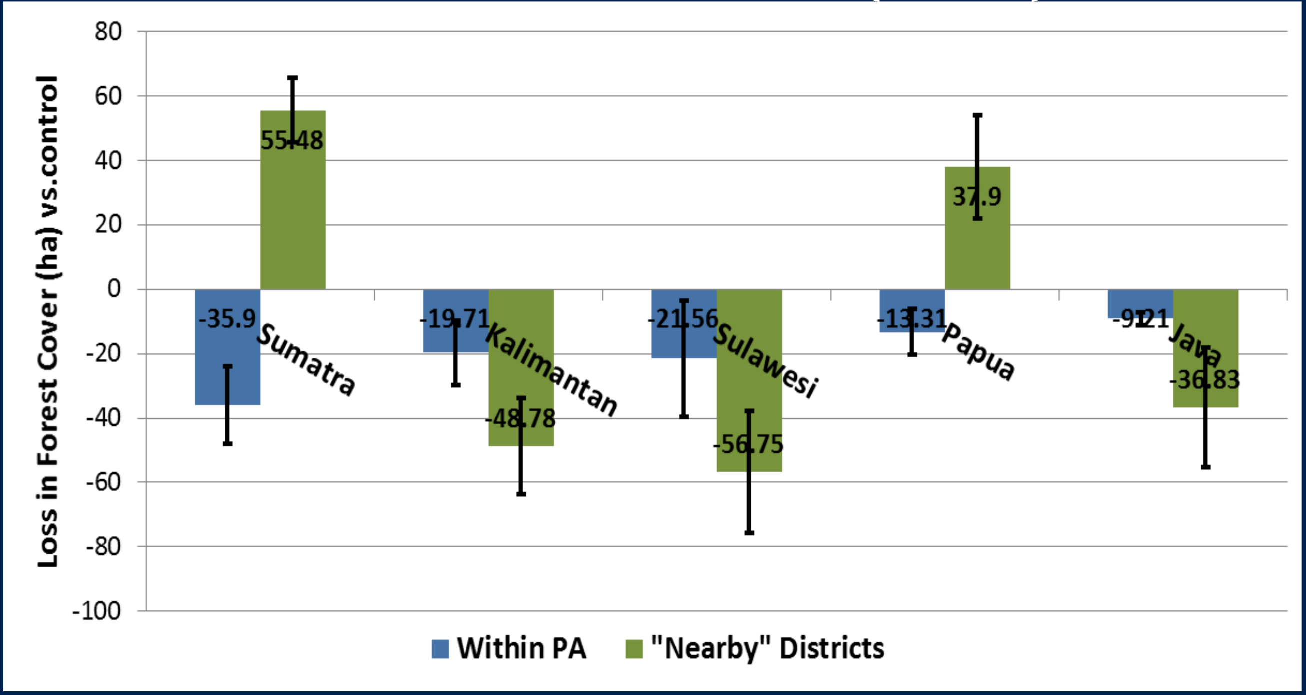
- Only some people eligible: compare ineligible people to controls



Between Villages: Even if one randomizes....

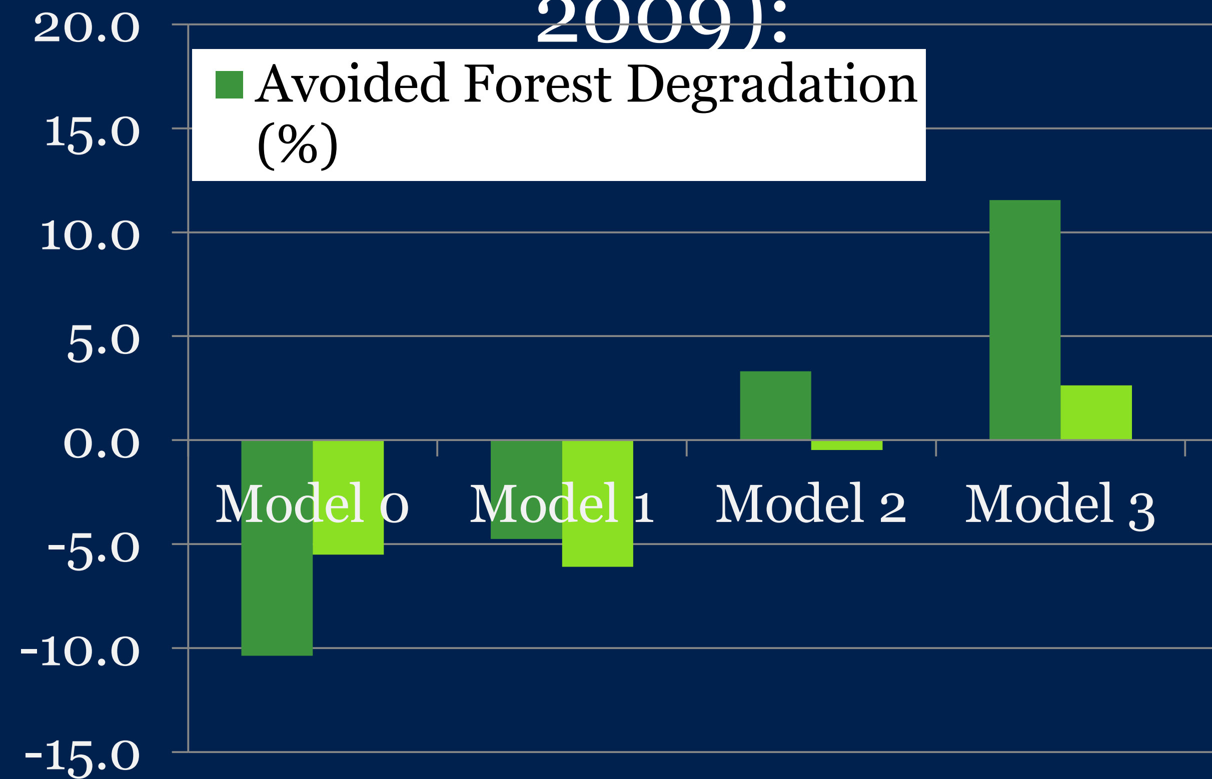
	Spatial Correlation Parameter					
	0.00	0.10	0.25	0.50	0.75	0.90
DD						
% Bias	-0.9	1.2	3.3	21.2	83.0	282.4
Rejection rate (95% Conf.)	93.4	94.5	92.1	86.0	68.1	50.2
DD with village fixed-effects						
% Bias	-0.9	1.2	3.3	21.2	83.0	282.4
Rejection rate (95% Conf.)	93.2	94.3	92.0	85.9	68.1	50.1
DD with individual fixed-effects						
% Bias	-0.9	1.2	3.3	21.2	83.0	282.4
Rejection rate (95% Conf.)	75.6	77.4	73.8	61.4	35.7	18.5
Spatial AR-DD						
% Bias	-0.9	0.7	-0.8	-0.2	0.3	0.2
Rejection rate (95% Conf.)	93.6	94.7	92.7	93.9	93.2	94.2

Spillovers: Forest Leakage from Protected Areas (PAs)



- Model 0: DiD, FE
- Model 1: DiD with Matching
- Model 2: DiD with Spatial Matching
- Model 3: Removing neighbouring controls

Avoided forest loss (1993 vs 2009):



Even without explicit spillovers…

- Error terms across neighbouring observations may be correlated
  - E.g. plot level data correlated by household
  - All households in a village being treated
  - Clustering standard errors

Spatially-correlated errors

	Spatial Correlation Parameter					
	0.00	0.10	0.25	0.50	0.75	0.90
DD						
% Bias	-0.1	-0.9	-0.6	-1.2	-0.6	4.4
Rejection rate (95% Conf.)	87.1	86.8	86.2	80.6	57.7	19.1
DD with village fixed-effects						
% Bias	-0.1	-0.9	-0.6	-1.2	-0.6	4.4
Rejection rate (95% Conf.)	87.0	86.5	86.0	80.6	57.7	19.1
DD with individual fixed-effects						
% Bias	-0.1	-0.9	-0.6	-1.2	-0.6	4.4
Rejection rate (95% Conf.)	64.6	65.3	62.6	54.1	25.4	4.9
Spatial Error-DD						
% Bias	-0.1	-0.9	-0.5	-1.1	-0.1	1.9
Rejection rate (95% Conf.)	87.7	86.6	87.1	83.0	78.6	76.2

Summary about SUTVA

- Set experimental design to minimize SUTVA

or…

- Build spillovers into the evaluation
- The spillovers may be interesting in and of themselves