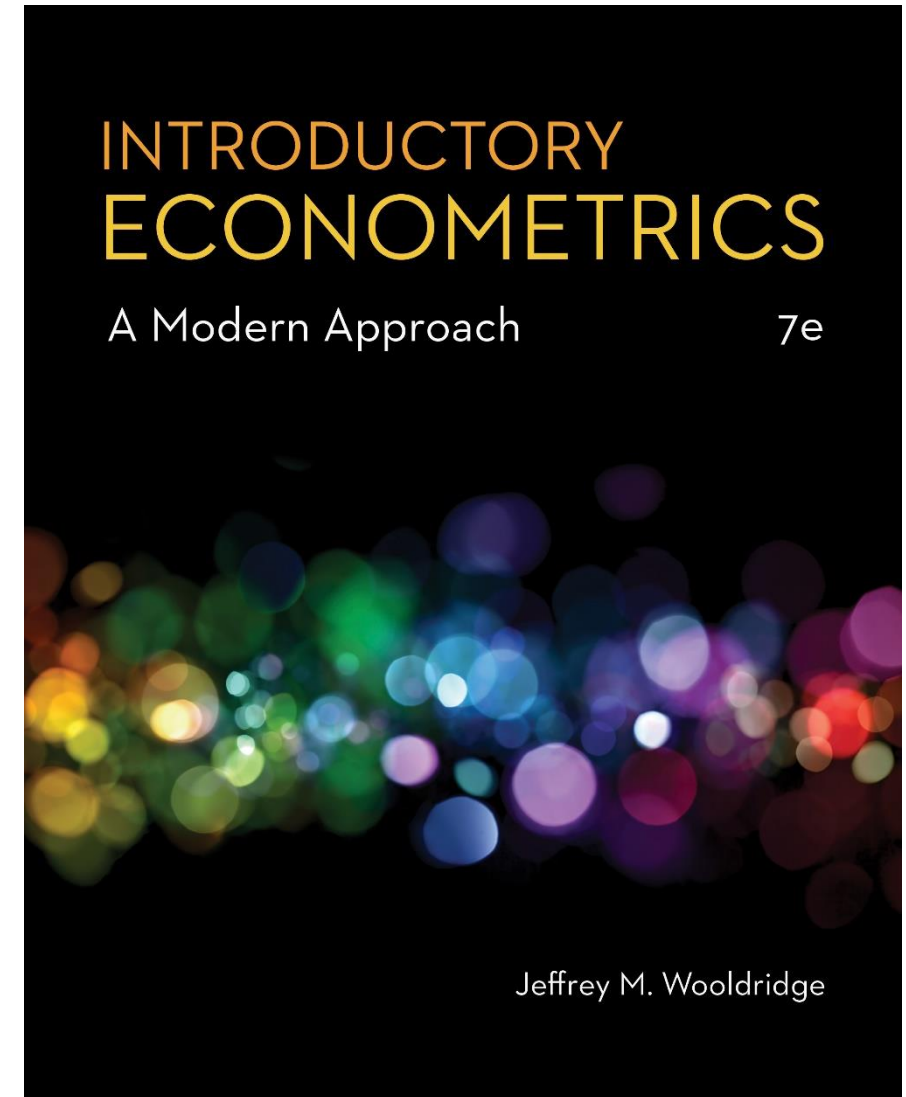


Chapter 13

Pooling Cross Sections across Time: Simple Panel Data Methods



Simple Panel Data Methods (1 of 13)

- **Policy analysis with pooled cross sections**
 - Two or more independently sampled cross sections can be used to evaluate the impact of a certain event or policy change.
 - Example: Effect of new garbage incinerator's location on housing prices.
 - Examine the effect of the location of a house on its price before and after the garbage incinerator was built:

$$\widehat{rprice} = 101,307.5 - 30,688.27 \text{ nearinc} \quad \leftarrow \text{After incinerator was built}$$

(3,093.0) (5,827.71)

$n = 142, R^2 = .165$

$$\widehat{rprice} = 82,517.23 - 18,824.37 \text{ nearinc} \quad \leftarrow \text{Before incinerator was built}$$

(2,653.79) (4,744.59)

$n = 179, R^2 = .082$

Simple Panel Data Methods (2 of 13)

• **Example: Garbage incinerator and housing prices (cont.)**

- It would be wrong to conclude from the regression after the incinerator is there that being near the incinerator depresses prices so strongly.
- One has to compare with the situation before the incinerator was built:

$$\hat{\delta}_1 = -30,688.27 - (-18,824.37) = -11,863.9$$



Incinerator depresses prices but location was one with lower prices anyway


- In the given case, this is equivalent to

$$\hat{\delta}_1 = (\overline{rprice}_{1,nr} - \overline{rprice}_{1,fr}) - (\overline{rprice}_{0,nr} - \overline{rprice}_{0,fr})$$

- This is called the difference-in-differences estimator (DiD)

Simple Panel Data Methods (3 of 13)

- **Difference-in-differences in a regression framework**

$$rprice = \beta_0 + \delta_0 after + \beta_1 nearinc + \delta_1 after \cdot nearinc + u$$


Differential effect of being in the location and after the incinerator was built

- In this way standard errors for the DiD-effect can be obtained.
 - If houses sold before and after the incinerator was built were systematically different, further explanatory variables should be included.
 - This will also reduce the error variance and thus standard errors.
- Before/After comparisons in “natural experiments”
 - DiD can be used to evaluate policy changes or other exogenous events.

Simple Panel Data Methods (4 of 13)

• Policy evaluation using difference-in-differences

$$y = \beta_0 + \delta_0 \text{after} + \beta_1 \text{treated} + \delta_1 \text{after} \cdot \text{treated} + \text{other factors}$$

$$\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{1,C}) - (\bar{y}_{0,T} - \bar{y}_{0,C}) \leftarrow \text{Compare outcomes of the two groups before and after the policy change}$$

- Compare the difference in outcomes of the units that are affected by the policy change (= treatment group) and those who are not affected (= control group) before and after the policy was enacted.
- For example, the level of unemployment benefits is cut but only for group A (= treatment group). Group A normally has longer unemployment durations than group B (= control group). If the difference in unemployment durations between group A and group B becomes smaller after the reform, reducing unemployment benefits reduces unemployment duration for those affected.
- Caution: Difference-in-differences only works if the difference in outcomes between the two groups is not changed by other factors than the policy change (e.g. there must be no differential trends).

Simple Panel Data Methods (5 of 13)

- **Two-period panel data analysis**
- Example: Effect of unemployment on city crime rate
 - Assume that no other explanatory variables are available. Will it be possible to estimate the causal effect of unemployment on crime?
 - Yes, if cities are observed for at least two periods and other factors affecting crime stay approximately constant over those periods:

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}, \quad t = 1982, 1987$$

Time dummy for
the second period

Unobserved time-constant
factors (= fixed effect)


Other unobserved factors
(= idiosyncratic error)

Simple Panel Data Methods (6 of 13)

- Example: Effect of unemployment on city crime rate (cont.)**

$$crmrte_{i1987} = \beta_0 + \delta_0 \cdot 1 + \beta_1 unem_{i1987} + a_i + u_{i1987}$$


$$crmrte_{i1982} = \beta_0 + \delta_0 \cdot 0 + \beta_1 unem_{i1982} + a_i + u_{i1982}$$

Subtract: $\Rightarrow \Delta crmrte_i = \delta_0 + \beta_1 \Delta unem_i + \Delta u_i$  Fixed effect drops out

- Estimate differenced equation by OLS:

$$\Delta \widehat{crmrte} = 15.40 + 2.22 \Delta unem$$

(4.70) (.88)

 + 1 percentage point unemployment rate leads to 2.22 more crimes per 1,000 people

$n = 46, R^2 = .127$  Secular increase in crime

Simple Panel Data Methods (7 of 13)

- **Discussion of first-differenced panel estimator**
- Further explanatory variables may be included in original equation.
- Note that there may be arbitrary correlation between the unobserved time-invariant characteristics and the included explanatory variables.
 - OLS in the original equation would therefore be inconsistent.
 - The first-differenced panel estimator is thus a way to consistently estimate causal effects in the presence of time-invariant endogeneity.
- For consistency, strict exogeneity has to hold in the original equation.
- First-differenced estimates will be imprecise if explanatory variables vary only little over time (no estimate possible if time-invariant).

Simple Panel Data Methods (8 of 13)

- **Another interpretation of the difference-in-differences estimator**

- We can re-write the DiD estimator as:

$$\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{0,T}) - (\bar{y}_{1,C} - \bar{y}_{0,C})$$

- The first term is the difference in means over time for the treated group
 - This would be a good estimator of the policy effect only if no external factors changed across the two time periods.
- The second term is the difference in means over time for the control group.
 - Subtracting off this term hopefully controls for any changes in external factors that are common to both the treated and control groups, which will be the case when we have random assignment.
 - In this case, the DiD estimator can be interpreted as the average treatment effect.

Simple Panel Data Methods (9 of 13)

- **Adding an additional control group**
- The standard two-group, two period difference-in-differences setup relies on the assumption of parallel trends.
 - Parallel trends assumes that any trends in the outcome y would trend at the same rate in the absence of the intervention.
 - Prior to the intervention, y should move in the same direction for both groups.
- The standard DiD estimator measures the difference in estimated trends between the two groups.
 - If the parallel trends assumption is violated, we cannot be sure that the DiD estimator is identifying the effects of the policy or simply some other unaccounted factor causing different trends between these groups.
- We can add flexibility by adding an additional control group.

Simple Panel Data Methods (10 of 13)

- **Adding an additional control group (contd)**
- Example: The effects of expanding health care for low income families in a particular state.
 - Let L denote low-income families (eligible for the policy) and M be middle-income families (not eligible).
 - Let B denote states that implemented the policy and A be states that did not implement the policy.
 - The policy is implemented in period 1, but no policy exists in period 0.
- The additional control group (income level) allows for more flexibility if we assume that any difference in trends in health outcomes between low and middle income families is similar across states.

Simple Panel Data Methods (11 of 13)

• Adding an additional control group (contd)

$$y = \beta_0 + \beta_1 dL + \beta_2 dB + \beta_3 dL * dB + \delta_0 d1 + \delta_1 d1 * dL + \delta_2 d1 * dB + \delta_3 d1 * dL * dB + u$$

$$\hat{\delta}_3 = [(\bar{y}_{1,L,B} - \bar{y}_{0,L,B}) - (\bar{y}_{1,M,B} - \bar{y}_{0,M,B})] - [(\bar{y}_{1,L,A} - \bar{y}_{0,L,A}) - (\bar{y}_{1,M,A} - \bar{y}_{0,M,A})]$$

$$\hat{\delta}_3 = \hat{\delta}_{DD,B} - \hat{\delta}_{DD,A} = \hat{\delta}_{DDD}$$

- The difference-in-difference-in differences estimator has two components
 - A DD estimator looking only at states that implemented the policy.
 - A DD estimator looking only at states that did not implement the policy.
- If health trends between the L and M groups do not differ in non-implementation states, then the second component vanishes and we are back to the standard DiD setup.
 - However, we include this second term to account for possibly different trends in the L and M groups that are common across both states A and B.

Simple Panel Data Methods (12 of 13)

- **A General Framework for Policy Analysis**
- A more general approach to policy analysis is to include multiple control and treatment groups as well as more than two time periods.
 - Some units may never be treated and others may be treated in different time periods.
 - With this general framework, we should avoid trying to fit the problem into the basic DD setup.
- Let each observation i belong to a pair (g, t) , where g is a group and t is a time period.
 - We are interested in policy interventions that occur at the group level and in order to be convincing, there should be before and after periods for at least some of the groups in our study.

Simple Panel Data Methods (13 of 13)

• A General Framework for Policy Analysis (contd)

- Indicate the policy with a dummy variable $x_{g,t}$ for group g and time t

$$y_{i,g,t} = \lambda_t + \alpha_g + \beta x_{g,t} + \mathbf{z}_{i,g,t}\boldsymbol{\gamma} + u_{i,g,t}$$

Aggregate time
effect common to
all groups

Fixed effects
specific to each
group

Policy effect

Explanatory variables
measured at the
individual and group level

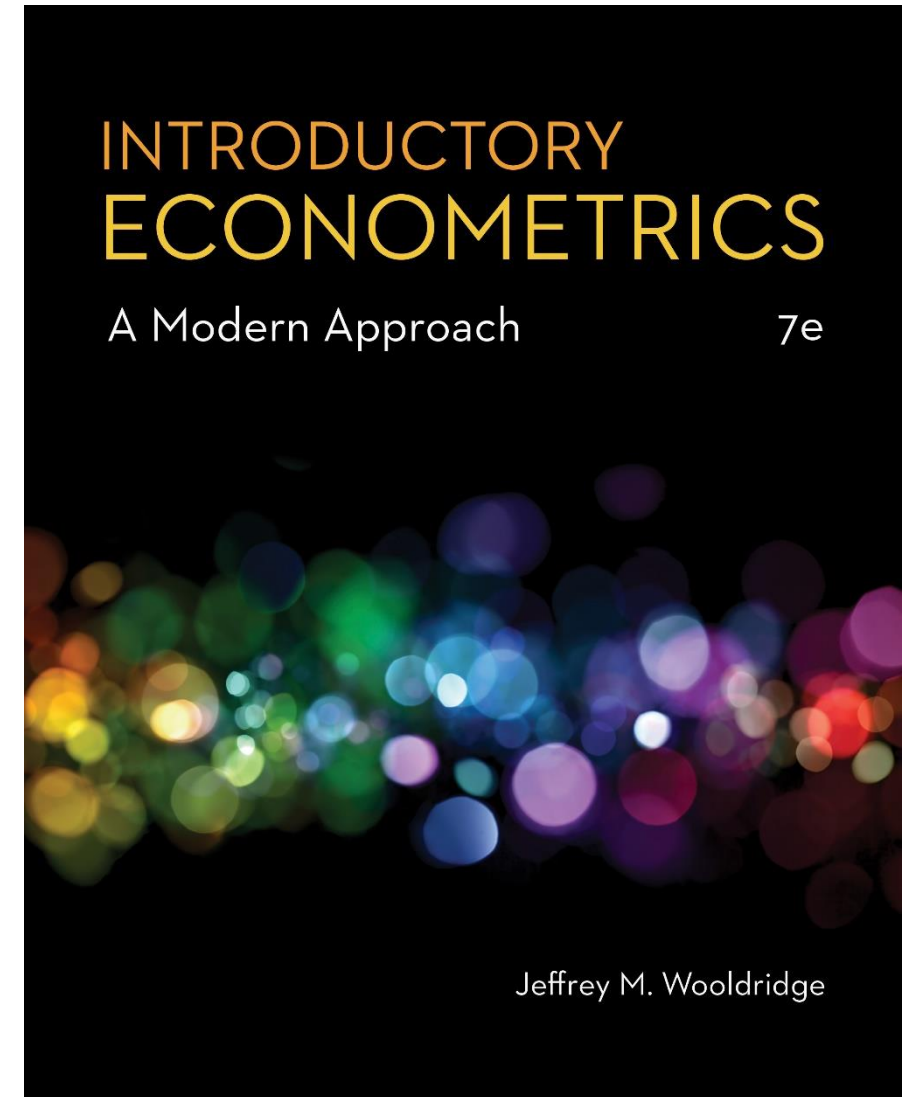
- This model can be expanded upon with the addition of lagged policy effects or allowing the policy to have different effects between groups.

- The parallel trends assumption can also be relaxed (for $T > 2$):

$$y_{i,g,t} = \lambda_t + \alpha_g + \psi_g t + \beta x_{g,t} + \mathbf{z}_{i,g,t}\boldsymbol{\gamma} + u_{i,g,t} \quad \leftarrow \psi_g \text{ captures the linear time trend by group}$$

Chapter 14

Advanced Panel Data Methods



Advanced Panel Data Methods (1 of 15)

• Fixed effects estimation

$$y_{it} = \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}, \quad i = 1, \dots, N, t = 1, \dots, T$$

Fixed effect, potentially correlated
with explanatory variables

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \dots + \beta_k \bar{x}_{ik} + \bar{a}_i + \bar{u}_i$$

Form time-averages
for each individual

$$\Rightarrow [y_{it} - \bar{y}_i] = \beta_1 [x_{it1} - \bar{x}_{i1}] + \dots + \beta_k [x_{itk} - \bar{x}_{ik}] + [u_{it} - \bar{u}_i]$$

Because $a_i - \bar{a}_i = 0$ (the fixed effect is removed)

- Estimate time-demeaned equation by OLS
 - Uses time variation within cross-sectional units (= within estimator)

Advanced Panel Data Methods (2 of 15)

• Example: Effect of training grants on firm scrap rate

$$scrap_{it} = \beta_1 d88_{it} + \beta_2 d89_{it} + \beta_3 grant_{it} + \beta_4 grant_{it-1} + a_i + u_{it}$$

Time-invariant reasons why one firm is more productive than another are controlled for.
The important point is that these may be correlated with the other explanatory variables.

• Fixed-effects estimation using the years 1987, 1988, and 1989:

$$\widehat{scrap}_{it}^* = - .080 \underset{(.109)}{d88_{it}^*} - .247 \underset{(.133)}{d89_{it}^*} - .252 \underset{(.151)}{grant_{it}^*} - .422 \underset{(.210)}{grant_{it-1}^*}$$

← Stars denote time-demeaning

$n = 162, R^2 = .201$ Training grants significantly improve productivity (with a time lag)

Advanced Panel Data Methods (3 of 15)

- **Discussion of fixed effects estimator**

- Strict exogeneity in the original model has to be assumed.
- The R-squared of the demeaned equation is inappropriate.
- The effect of time-invariant variables cannot be estimated.
- The effect of interactions with time-invariant variables can be estimated (e.g. the interaction of education with time dummies).
- If a full set of time dummies are included, the effect of variables whose change over time is constant cannot be estimated (e.g. experience).
- Degrees of freedom have to be adjusted because the N time averages are estimated in addition (resulting degrees of freedom = $NT - N - k$).

Advanced Panel Data Methods (4 of 15)

- **Interpretation of fixed effects as dummy variable regression**

- The fixed effects estimator is equivalent to introducing a dummy for each individual in the original regression and using pooled OLS:

$$y_{it} = a_1 ind1_{it} + a_2 ind2_{it} + \dots + a_N indN_{it} \leftarrow \text{For example, } = 1 \text{ if the observation stems from individual } N, = 0 \text{ otherwise}$$

$$+ \beta_1 x_{it1} + \dots + \beta_k x_{itk} + u_{it}$$

- After fixed effects estimation, the fixed effects can be estimated as:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \dots - \hat{\beta}_k \bar{x}_{ik}, \quad i = 1, \dots, N \leftarrow \text{Estimated individual effect for individual } i$$

Advanced Panel Data Methods (5 of 15)

- **Fixed effects or first differencing?**
 - Remember that first differencing can also be used if $T > 2$.
 - In the case $T = 2$, fixed effects and first differencing are identical.
 - For $T > 2$, fixed effects is more efficient if classical assumptions hold.
 - First differencing may be better in the case of severe serial correlation in the errors, for example if the errors follow a random walk.
 - If T is very large (and N not so large), the panel has a pronounced time series character and problems such as strong dependence arise.
 - In these cases, it is probably better to use first differencing.
 - Otherwise, it is a good idea to compute both and check robustness.

Advanced Panel Data Methods (6 of 15)

- **Unbalanced panels**

- An unbalanced panel is when not all cross-sectional units have the same number of observations.
- Dropping units with only one time period does not cause bias or inconsistency.

- **Fixed effects (FE) or First Differencing (FD) with unbalanced panels**

- FE will preserve more data than FD when we have unbalanced panels, since FD requires that each observation have data available for both t and $t-1$.
- For example, consider a scenario in which we have seven years of data, but data is missing for all even numbered years. Thus, we observe $t=1,3,5,7$.
 - FE will use time periods 1,3,5,7
 - FD will lose all observations.

Advanced Panel Data Methods (7 of 15)

• Random effects (RE) models

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}$$

← The individual effect is assumed to be “random”
i.e. completely unrelated to explanatory variables

Random effects assumption: $Cov(x_{itj}, a_i) = 0, j = 1, 2, \dots, k$

- The composite error $a_i + u_{it}$ is uncorrelated with the explanatory variables but it is serially correlated for observations coming from the same i :

$$Cov(a_i + u_{it}, a_i + u_{is}) = Cov(a_i, a_i) = \sigma_a^2$$

← Under the assumption that
idiosyncratic errors are
serially uncorrelated

For example, in a wage equation, for a given individual the same unobserved ability appears in the error term of each period. Error terms are thus correlated across periods for this individual.

Advanced Panel Data Methods (8 of 15)

• Estimation in the random effects model

- Under the random effects assumptions explanatory variables are exogenous so that pooled OLS provides consistent estimates.
- If OLS is used, standard errors have to be adjusted for the fact that errors are correlated over time for given i (= clustered standard errors).
- But, because of the serial correlation, OLS is not efficient.
- One can transform the model so that it satisfies the GM-assumptions:

$$[y_{it} - \lambda \bar{y}_i] = \beta_1 [x_{it1} - \lambda \bar{x}_{i1}] + \dots + \beta_k [x_{itk} - \lambda \bar{x}_{ik}] \leftarrow \text{Quasi-demeaned data}$$

$$+ [a_i - \lambda \bar{a}_i + u_{it} - \lambda \bar{u}_i] \leftarrow \text{Error can be shown to satisfy GM-assumptions}$$

Advanced Panel Data Methods (9 of 15)

- **Estimation in the random effects model (cont.)**

$$\lambda = 1 - \left[\sigma_u^2 / (\sigma_u^2 + T\sigma_a^2) \right]^{1/2}, \quad 0 \leq \lambda \leq 1$$

- The quasi-demeaning parameter is unknown but it can be estimated.
- FGLS using the estimated λ is called random effects estimation.
- If the random effect is relatively unimportant compared to the idiosyncratic error, FGLS will be close to pooled OLS (because λ goes to 0).
- If the random effect is relatively important compared to the idiosyncratic term, FGLS will be similar to fixed effects (because λ goes to 1).
- Random effects estimation can be used to estimate the effect of time-invariant variables.

Advanced Panel Data Methods (10 of 15)

• Example: Wage equation using panel data

$$\begin{aligned}\widehat{\log(wage_{it})} = & \underset{(.011)}{.092} educ_{it} - \underset{(.048)}{.139} black_{it} + \underset{(.043)}{.022} hispan_{it} \\ & + \underset{(.015)}{.106} exper_{it} - \underset{(.0007)}{.0047} exper_{it}^2 + \underset{(.017)}{.064} married_{it} \\ & + \underset{(.018)}{.106} union_{it} + time\ dummies\end{aligned}$$

Random effects is used because many of the variables are time-invariant. But is the random effects assumption realistic?

• Random effects or fixed effects?

- In economics, unobserved individual effects are seldomly uncorrelated with explanatory variables so that fixed effects is more convincing.

Advanced Panel Data Methods (11 of 15)

• Correlated Random Effects (CRE)

- When using CRE to choose between FE and RE, we must include any time-constant variables that appear in RE estimation:

$$y_{it} = \alpha_1 + \alpha_2 d2_t + \cdots + \alpha_T dT_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} \\ + \gamma_1 \bar{x}_{i1} + \cdots + \gamma_k \bar{x}_{ik} + \delta_1 z_{i1} + \cdots + \delta_m z_{im} + r_i + u_{it}$$

- Estimating this equation by RE (or even just pooled OLS) yields:

$$\hat{\beta}_{CRE,j} = \hat{\beta}_{FE,j}; j = 1, \dots, k \quad \leftarrow \text{Time varying estimates will be the same as in FE}$$

$$\hat{\alpha}_{CRE,t} = \hat{\alpha}_{FE,t}; t = 1, \dots, T$$

$$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0 \quad \leftarrow \text{Under the null, RE is sufficient. If we reject the null, then FE is preferred.}$$

- An advantage of CRE is that it allows for estimation of the effects of time-constant explanatory variables, not possible using FE.

Advanced Panel Data Methods (12 of 15)

- **General policy analysis with panel data**

- The two-period, before-after setting is a special case of a more general policy analysis framework when $T \geq 2$.

$$y_{it} = \eta_1 + \alpha_2 d2_t + \cdots + \alpha_T dT_t + \beta w_{it} + \mathbf{x}_{it}\boldsymbol{\psi} + a_i + u_{it}$$

w_{it} is the binary policy variable and β estimates the average treatment effect of the policy

- To allow w_{it} to be systematically related to the unobserved fixed effect a_i , we estimate the regression with either FD or FE, using cluster-robust standard errors.
- We can also include lags of the policy intervention: $w_{it-1}, w_{it-2}, \dots$

Advanced Panel Data Methods (13 of 15)

- **Testing for feedback from the error term to the policy variable**
- We need to be careful if the policy variable w_{it} it reacts to past shocks.
 - Example: y_{it} is the poverty rate and w_{it} is some measure of government assistance.
 - A large shock to the poverty rate in year t could prompt an increase in government assistance the following year.

- If we have at least three time periods, we can test for feedback

$$y_{it} = \eta_1 + \alpha_2 d2_t + \cdots + \alpha_{T-1} dT - 1_t + \beta w_{it} + \delta w_{it+1} + x_{it} \psi + a_i + u_{it}$$

← Estimate with FE and compute a cluster robust t-statistic for $\hat{\delta}$

- This is known as a “falsification test.”
 - If the forward policy variable is statistically significant, there is potential feedback from the error term to the policy variable.

Advanced Panel Data Methods (14 of 15)

- **The heterogeneous trend model**
- What if time trends are unique across individuals?

$$y_{it} = \eta_1 + \alpha_2 d2_t + \cdots + \alpha_T dT_t + \beta w_{it} + \mathbf{x}_{it}\boldsymbol{\psi} + a_i + g_i t + u_{it}$$

The new term $g_i t$ is a unit-specific time trend.

- This allows the policy intervention to not only be correlated with level differences among units (captured by a_i), but also by trend differences.
- We can estimate this model by taking first differences:

$$\Delta y_{it} = \alpha_2 \Delta d2_t + \cdots + \alpha_T \Delta dT_t + \beta \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\psi} + g_i + \Delta u_{it}$$

Estimate by FE., though we need to ensure we have $T \geq 3$

Advanced Panel Data Methods (15 of 15)

- **Applying panel data methods to other data structures**
 - Panel data methods can be used in other contexts where constant unobserved effects have to be removed.
- Example: Wage equations for twins

Unobserved genetic and family characteristics that do not vary across twins

$$\log(wage_{i1}) = \beta_0 + \beta_1 educ_{i1} + \dots + a_i + u_{i1} \leftarrow \text{Equation for twin 1 in family } i$$

$$\log(wage_{i2}) = \beta_0 + \beta_1 educ_{i2} + \dots + a_i + u_{i2} \leftarrow \text{Equation for twin 2 in family } i$$

$$\Rightarrow \Delta \log(wage_i) = \beta_1 \Delta educ_i + \dots + \Delta u_i \leftarrow \text{Estimate differenced equation by OLS}$$