Lecture 11: Panel Data (Part I)

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

A Brief Introduction to Panel Data

- For the most part, we have focused on cross-sectional applications: random sample of N individuals at a single point in time
- We will now move to focus explicitly on variation over time in panel data
  - The same units are observed at multiple points in time
  - N individuals observed for T time periods
- You have probably already seen some examples of this!

Plan for today

1. Introduction
2. Pooling repeated cross-sections
3. Difference-in-differences (DiD)
4. Issues in pooled OLS and DiD
   - Is "parallel trends" a reasonable assumption?
   - Group-specific linear trends
   - Example: Sulfa drugs
   - Composition effects
5. First-differenced estimator
6. Fixed effects estimation

Pooling Independent Cross-Sections

- Our goal will be to estimate parameters identified using variation over time for the same population
- This is a very common case in practice and central to the econometric toolkit for applied economists
- The first and important case is where we have data on repeated cross-sections
  - Imagine that we draw a random sample of N individuals drawn at time t
  - In time $t+1$, we draw another random sample from the same underlying population
- As we will see, this analysis does not require much more by way of econometric tools than you have seen already—but the applications are very wide-ranging

Pooling Independent Cross-Sections: Have Parameters Changed over Time?

- Frequently, we are interested in whether specific relationships between variables have changed over time
- Imagine we are interested in whether the gender gap in test scores has changed over time in Sweden
- Suppose we have access to the PISA dataset from two rounds (say, 2012 and 2015)
- The easiest way to do this is to "pool" the cross-sections (in practice: append the datasets) and run a single regression

$$Y_{it} = \alpha_1 + \alpha_2 y2015_{it} + \beta_1 male + \beta_2 (male_i \times y2015_{it}) + \epsilon_{it}$$

where Y_{it} is the test score of child i at time t, *male* is a dummy variable equal to 1 if the child is a boy, and $y2015$ is a dummy variable equal to 1 if the observation is from 2015

Pooling Independent Cross-Sections: Incorporating Other Variables

- Suppose we also had other variables available—how would we include them in the analysis?
- Imagine we have an additional variable measuring family wealth
- Option 1: add as a separate control

$$\begin{aligned} Y_{it} &= \alpha_1 + \alpha_2 y2015_{it} + \beta_1 male + \beta_2 (male_i \times y2015_{it}) \\ &\quad + \gamma_1 wi_{it} + \epsilon_{it} \end{aligned}$$

- Option 2: include as control AND interact with year

$$\begin{aligned} Y_{it} &= \alpha_1 + \alpha_2 y2015_{it} + \beta_1 male + \beta_2 (male_i \times y2015_{it}) \\ &\quad + \gamma_1 wi_{it} + \gamma_2 (wi \times y2015_{it}) + \epsilon_{it} \end{aligned}$$

Pooling Independent Cross-Sections: Incorporating Other Variables

- Option 2 on the previous slide nests option 1
  - The two are the same if $\gamma_2 = 0$
  - You could also imagine a triple-interaction between the three variables
  - In practice, you will often see specifications like option 1
- The more general specification is to run the regressions separately for each year
- This allows the coefficients on all variables to differ across year
- Identical to estimating a regression which allows for interaction between year and all the RHS variables

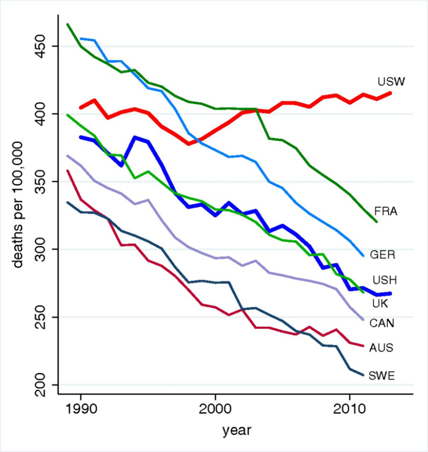
Policy Analysis with Pooled Cross-Sections

- Our core interest in this course has been on the estimation of causal effects
- One of the most common applications of such methods is where a policy varies over time
- Imagine, for instance, two cross-sections of the population
  - A pre-treatment period before any policy change has happened
  - A post-treatment period where some individuals (treatment) have been affected by a policy and others have not (control)
- The canonical method for estimating the effect of the treatment is the **difference-in-differences estimator**

Pooled OLS

- The descriptive example we discussed was one of just seeing how some parameters had changed over time
- In a broad class of questions we want to see, descriptively, whether gaps between groups have changed over time
- Pooled OLS, or equivalent methods, would be a standard way of analyzing and presenting these changes
- One **major** issue to keep in mind when doing this is composition effects
  - Have characteristics of individuals within the group changed?
  - This also affects DiD with pooled cross sections (and matters for identification)!

An Example of Composition Effects: The Rising Mortality of White Non-Hispanic Americans (Case and Deaton 2015)



But Perhaps It Is Only Rising Because of Composition Effects... (Gelman and Auerbach 2016)

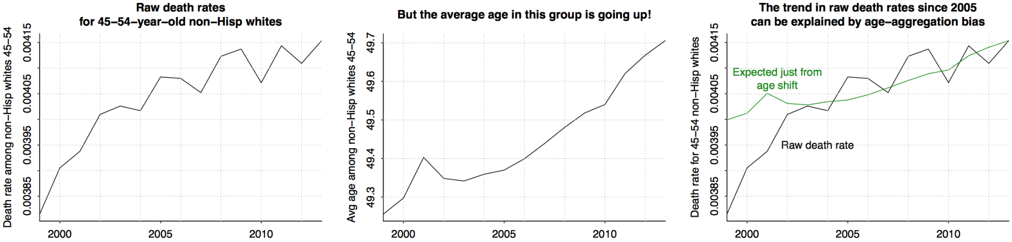


Fig. 1. (a) Observed increase in raw mortality rate among 45-54-year-old non-Hispanic whites, unadjusted for age; (b) Increase in average age of this group as the baby boom generation moves through; (c) Raw death rate, along with trend in death rate attributable by change in age distribution alone, had age-specific mortality rates been at the 2013 level throughout.

Composition Effects

- Note that the issue of composition effects is particularly tricky for working with repeated cross-sections
  - The underlying population could be changing
  - With individual-level panel data we track the same individuals over time
- This issue is central to not just descriptives but also to assess identification in DiD-type specifications
- If the treatment population is changing differently to untreated population over time, we have a violation of the parallel trends assumption!

The Difference-in-Differences Estimator

- The difference-in-differences (DiD) estimator, with two periods, is similar to the specification we saw before:

$$Y_{igt} = \alpha_1 + \alpha_2 POST_t + \beta_1 Treat_g + \beta_2 (Treat_g \times POST_t) + \gamma X_{igt} + \epsilon_{igt}$$

  where i indexes individuals, g indexes groups, t indexes time, $POST$ is a dummy variable for being in the post-treatment period, $Treat$ is a dummy variable for being in the treatment group, and X is a vector of other variables (assumed to have the same coefficient in both periods; could be generalized)

The Difference-in-Differences Estimator

- Imagine, for instance, wanting to study the effect of minimum wages on employment
  - Two states, A and B
  - Two time periods, 1 and 2
  - State A introduces a minimum wage in period 2

$$Y_{igt} = \alpha_1 + \alpha_2 POST_t + \beta_1 Treat_g + \beta_2 (Treat_g \times POST_t) + \epsilon_{igt}$$

- Ignoring covariates, we have the following outcomes:

	State A (T)	State B (C)
Period 1 (pre)	$\alpha_1 + \beta_1$	α_1
Period 2 (post)	$\alpha_1 + \alpha_2 + \beta_1 + \beta_2$	$\alpha_1 + \alpha_2$

The Difference-in-Differences Estimator

$$Y_{igt} = \alpha_1 + \alpha_2 POST_t + \beta_1 Treat_g + \beta_2(Treat_g \times POST_t) + \epsilon_{igt}$$

$$\widehat{\beta_2} = (\bar{Y}_{T,2} - \bar{Y}_{C,2}) - (\bar{Y}_{T,1} - \bar{Y}_{C,1})$$

- **Before-after estimator:** $(\bar{Y}_{T,2} - \bar{Y}_{T,1}) = \alpha_2 + \beta_2$ (ignores that outcomes may have secular change over time)
- **Treatment-control estimator:**$(\bar{Y}_{T,2} - \bar{Y}_{C,2}) = \beta_1 + \beta_2$ (ignores that outcomes may have differed between T and C even without treatment)
- **Difference-in-differences estimator:** $(\bar{Y}_{T,2} - \bar{Y}_{C,2}) - (\bar{Y}_{T,1} - \bar{Y}_{C,1})$ (deals with levels differences between groups and between years)

The Difference-in-Differences Estimator (Angrist and Pischke 2009)

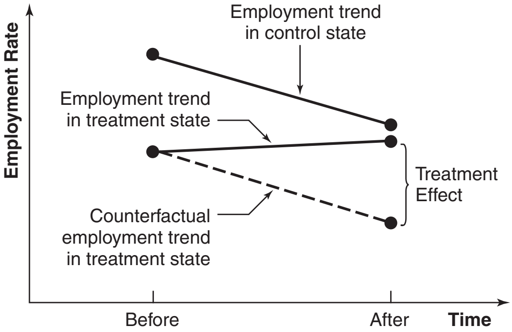


Figure 5.2.1 Causal effects in the DD model.

The Difference-in-Differences Estimator: Adding Covariates

- Incorporating covariates is easy:

$$Y_{igt} = \alpha_1 + \alpha_2 POST_t + \beta_1 Treat_g + \beta_2 (Treat_g \times POST_t) + \gamma X_{igt} + \epsilon_{igt}$$

- When other explanatory variables are added in, we would still interpret β_2 as the DiD estimator
- But note that now $\beta_2 \neq (\bar{Y}_{T,2} - \bar{Y}_{C,2}) - (\bar{Y}_{T,1} - \bar{Y}_{C,1})$
  - The intuitive way to think about β_2 now is that we have first partialled out the effect of the other variables
  - And then we are looking at the difference-in-differences as earlier

The Difference-in-Differences Estimator: Parallel Trends

- The causal interpretation of the DiD estimator requires that (potentially conditional on X) the treatment and control groups **would have had similar changes in outcomes in the absence of treatment**
- This is the **"parallel trends" assumption**
- What this implies:
  - The treatment and control groups **can** have different levels of Y...
  - ...but not different growth rates in Y

The Difference-in-Differences Estimator: Extensions

- It is easy to allow different effects for different time periods instead of comparing just pre versus post
  - As we will see, this can sometimes be quite useful for substantiating the parallel trends assumption!
- Treatment does not have to happen at the same time for all treated units—it could happen in a staggered manner
  - We could then estimate something called a two-way fixed effects specification
  - Or we could estimate an event-study specification
  - Caveat: staggered treatments also involve complicated econometric issues that we will not be able to discuss in this class
- We can also have a continuous treatment instead of the binary one (interpretation slightly trickier, though)
- We will circle back to these issues later (after having discussed fixed effects)—before that, let us consider other issues in pooled OLS and DiD...

Evidence in Support of Parallel Trends

- Note the identifying assumption in DiD is that trends would have been similar across treated and untreated groups, in the absence of treatment
  - This is fundamentally unverifiable
  - Y_0 is not observed for treated units post-treatment!

- But there are several things that you can do in order to look at whether parallel trends is a reasonable assumption

- With ≥ 2 rounds of pre-treatment data, you can test whether trends were parallel before the policy was introduced

- This could be done formally and also presented graphically

Group-Specific (Linear) Trends

- In the canonical DiD model set-up, there are two groups and two time periods (treated and control + pre and post)

- In that set-up, there is little you can do to guard against a violation of parallel trends...

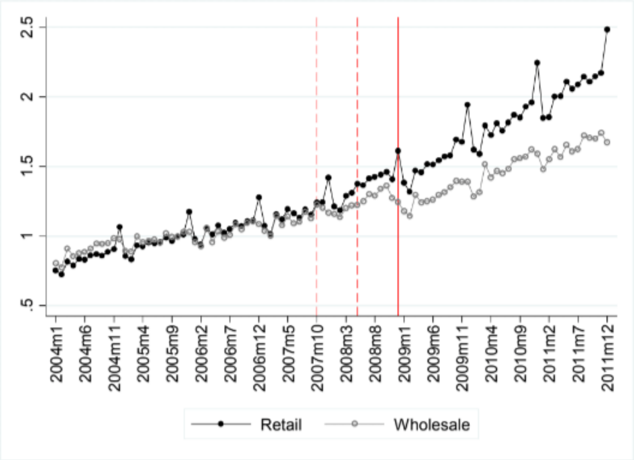
- But often you do have two or more pre-treatment periods

- If you are worried that parallel trends do not seem plausible (and even if not), you could also include linear (or some other) time trends for the different groups

$$Y_{it} = \alpha_0 + \alpha_1 Treat_i + \alpha_2 time_t + \alpha_3 (Treat_i \times time_t) + \alpha_4 (POST_t \times Treat_i) + \epsilon_{it}$$

where *time* is a linear time trend, *POST* is a post-treatment indicator, *Treat* is a treatment indicator

Graphical Evidence in Support of Parallel Trends (Naritomi 2018)



a. Raw data: reported revenue changes

Evidence in Support of Parallel Trends

- Suppose you have data on units that belong to a treatment group and a control group
- Suppose we observe the outcomes for both groups at three points of time, $t = -1$, $t = 0$, and $t = 1$
  - $t = -1$: treatment has not happened yet
  - $t = 0$: outcomes are measured at the beginning of this time period, after which the treatment takes place
  - $t = 1$: treated units have been treated
- We could run the following regression to understand whether the trends between the control and the treatment group differed before the treatment:

$$Y_{it} = \alpha + \beta 1[t = -1]_t + \gamma 1[t = 1]_t + \delta T_i + \theta 1[t = -1]_t T_i + \rho 1[t = 1]_t T_i + \epsilon_{it}$$

- θ tells us the difference-in-differences between the treatment and the control units between $t = -1$ and $t = 0$ and ρ is the difference-in-differences estimate of the treatment effect

Evidence in Support of Parallel Trends

- With just three periods of data, you could also look at the change between $t = -1$ and $t = 0$ as the dependent variable

- But you might also have more data—then you can easily extend this to multiple time periods

- Remember to leave one of the time periods as the baseline (say, S, which could be the last point of measurement before a treatment takes place)!

$$Y_{it} = \alpha + \beta T_i + \sum_{s \neq S}^{T} (\gamma_s 1[t = s]_t + \delta_s 1[t = s]_t T_i) + \epsilon_{it}$$

- There is evidence of no pre-treatment trends if $\delta_s = 0$ for $s < S$

Making Use of Many Time Periods

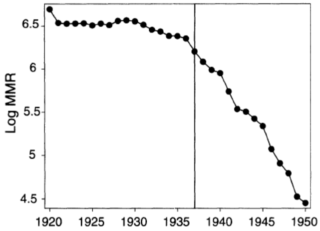
- The DiD model is very general and is easily modified to incorporate many time periods
- Multiple pre-treatment periods allow us to probe the common/parallel trends assumption
  - We can check whether the trends were similar before treatment
  - Major remaining worry: did anything else change at the same time?
  - One solution: we can allow for the trends to be different in the post-treatment period
- This is easiest to see using an example
- We will look at Jayachandran and Lleras-Muney (2010) which looks at the effect of a new type of antibiotic ("sulfa drugs") on mortality in the US

Modern Medicine and the Decline in 20th-Century Mortality (Jayachandran and Lleras-Muney 2010)

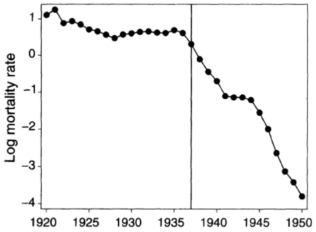
- Mortality dropped sharply in the 20th century
- Many potential causes:
  - Better water supply, sanitation and hygiene
  - Higher incomes and better nutrition
  - More effective medicines
- So it is hard to quantify the contribution of modern medicine to falling mortality in the first half of the 20th century
- Jayachandran & Lleras-Muney (2010) look at the introduction of sulfa drugs in the 1930s
  - First treatment for some bacterial infections (sepsis, pneumonia, scarlet fever)
  - Did not treat major other diseases (tuberculosis)

Mortality for Treated Diseases (Jayachandran and Lleras-Muney 2010)

Panel A. Log maternal mortality ratio (deaths per 100,000 live births)



Panel C. Log scarlet fever mortality rate per 100,000



Mortality for Control Diseases (Jayachandran and Lleras-Muney 2010)

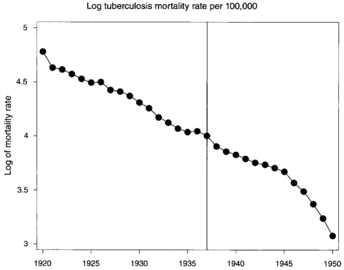


FIGURE 3. MORTALITY TRENDS (*in logs*) FOR CONTROL DISEASE

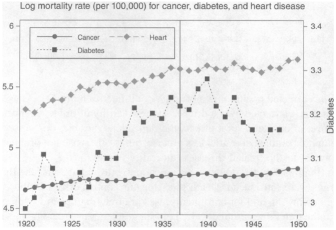


FIGURE 4. MORTALITY TRENDS (*in logs*) FOR CHRONIC DISEASES

Difference-in-Differences Specifications (Jayachandran and Lleras-Muney 2010)

$$(5) \quad \log(M)_{dt} = \beta_0 + \beta_1 \, \mathit{Treated}_d \times \mathit{Post\text{-}1937}_t + \beta_2 \, \mathit{Treated}_d \times \mathit{Year}_t + \beta_3 \, \mathit{Treated}_d + \beta_4 \, \mathit{Year}_t + \beta_5 \, \mathit{Post\text{-}1937}_t + \varepsilon_{dt}$$

and

$$(6) \quad \log(M)_{dt} = \beta_0 + \beta_1 \, \mathit{Treated}_d \times \mathit{Year}_t \times \mathit{Post\text{-}1937}_t + \beta_2 \, \mathit{Treated}_d \times \mathit{Post\text{-}1937}_t + \beta_3 \, \mathit{Treated}_d \times \mathit{Year}_t + \beta_4 \, \mathit{Treated}_d + \beta_5 \mathit{Year}_t + \beta_6 \, \mathit{Post\text{-}1937}_t + \varepsilon_{dt}.$$

Results (Jayachandran and Lleras-Muney 2010)

TABLE 4—EFFECTS OF SULFA DRUGS ON MORTALITY FOR "TREATED" DISEASES, 1937–1943

Dependent variable = ln (mortality)	MMR		Pneumonia/influenza		Scarlet fever	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel A. National-level data, all years, 1925–1943						
Treated × post-1937	−0.319**	−0.163***	−0.178	−0.052	−0.877**	−0.510***
	(0.118)	(0.041)	(0.176)	(0.126)	(0.337)	(0.110)
Treated × year × post-1937		−0.108***		−0.087***		−0.254***
		(0.009)		(0.031)		(0.036)
Observations	38	38	38	38	38	38
R^2	0.99	1.00	0.91	0.95	0.99	1.00

Individual-Level Panel Data Analysis

- Imagine that we have individual-level panel data
- Assume we have two periods, $t = 1$ and $t = 2$, for which each individual is observed
- Omitted variable bias is a persistent problem in a cross-sectional setting
- But with panel data, we can control for more and more factors, but some factors cannot be observed (e.g., ability in an earnings regression)...
- With two periods, “first-differencing” could be a solution!

Time-invariant Unobservables: A Two-Period Setting

- We can distinguish unobservables that are constant over time and those that vary over time:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$$

- a_i **captures all unobserved, time-constant factors affecting** y_{it}
- u_{it} captures time-varying unobserved factors
- i denotes the cross-sectional unit and t denotes time
- $d2_t$ is a dummy for the second period

"Naïve" Pooled OLS

- We could pool the two years and use OLS
- The model we estimate is:
$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}$$
with composite error $v_{it} = a_i + u_{it}$
- This will give us a consistent estimator if a_i is uncorrelated with x_{it}
- But... **OLS is inconsistent** if v_{it} is correlated with x_{it}, which would be the case **if a_i is correlated with x_{it}**

We Can Elimate a_i by Differencing

- Write down the model for both periods

$$y_{i2} = \beta_0 + \delta_0 + \beta_1 x_{i2} + a_i + u_{i2} \text{ for } t = 2$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \text{ for } t = 1$$

- Subtract the second from the first equation:

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

or

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- a_i has been differenced out, while the intercept in the differenced regression (δ_0) is the change in the intercept from $t = 1$ to $t = 2$

The First-Differenced Estimator

- Applying OLS to this model gives us the **first-differenced estimator**
  - For this estimator to be consistent, we need that Δu_i **is uncorrelated with** Δx_i
  - This will be the case if u_{it} in each period is uncorrelated with x_{it} in both periods (strict exogeneity)
  - Also, this rules out cases where y_{it-1} is used as a covariate (more on this later)

- The FD estimator is not restricted to two periods; we could have T time periods and run an FD specification

Differencing Has Some Drawbacks...

1 If the key covariates do not change much over time, **differencing will get rid of most of the information** in these variables

- $\Rightarrow$ Imprecise estimators
- $\Rightarrow$ If a covariate does not change at all, we cannot even estimate its effect in a differenced model!
- $\Rightarrow$ More subtle problem: the **effect** that you identify might **not really** be **relevant for the population at large**

Example

- Consider estimating the rate of return to education:

$$\ln(wage_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}$$

  where a_i contains ability

- Differencing the model

$$\Delta \ln(wage_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i$$

  gets rids of ability

- But since we are interested in working individuals the change in education will also be zero for most observations

- Thus β_1 will be identified off the back of a small number of people who get more education while they are already working

Drawbacks from Differencing

2 Differencing might also **make problems of measurement error worse**

- If the true variation in x across time periods is small, while measurement error is less persistent, Δx will consist mainly of measurement error

$\Rightarrow$ This makes the measurement error problem more severe

- Why? x contains two parts: a true part (the correctly measure variable) plus some measurement error
- Often, this measurement error leads to a correlation between the measured variable and the error term
- If the measurement error is not persistent but the true variation is, then differencing will only emphasize the measurement error!

Also the "Within-Transformation" Gets Rid of a_i

- Suppose we have multiple periods of data and we are interested in the following model

$$y_{it} = \beta_1 x_{it} + a_i + u_{it} \text{ for } t = 1, 2, \ldots, T$$

- For each i, average this equation over time:

$$\overline{y}_i = \beta_1 \overline{x}_i + a_i + \overline{u}_i$$

  where $\overline{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ and similar for $\overline{x}_i$ and $\overline{u}_i$

- Since a_i is fixed $\overline{a}_i = T^{-1} \sum_{t=1}^{T} a_i = a_i$

Also the "Within-Transformation" Gets Rid of a_i

- If we subtract the equation containing the time-averaged variables from the original equation, we get:

$$y_{it} - \overline{y}_i = \beta_1 \left(x_{it} - \overline{x}_i \right) + u_{it} - \overline{u}_i$$

which we can write as

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}$$

where $\ddot{y}_{it} = y_{it} - \overline{y}_i$ is the time-demeaned data for y, and similar for $\ddot{x}_{it}$ and $\ddot{u}_{it}$

Also the “Within-Transformation” Gets Rid of a_i

- **In the transformed equation a_i is gone!**
- Estimating the transformed model gives us the **fixed effects estimator** (or the within-estimator)
- “Within-estimator” because it uses the within variation in the data (i.e., the variation within each i across time) to identify the parameter of interest

Additional Covariates

- Nothing changes if we add additional covariates:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it} \text{ for } t = 1, 2, \ldots, T$$

- We just apply the time-demeaning to each explanatory variable, which yields

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + \ldots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it} \text{ for } t = 1, 2, \ldots, T$$

  and estimate this model with OLS

Fixed Effects (FE)

- Notice that, for FE to be unbiased and consistent, we need the x_{itj} to be strictly exogenous with respect to u_{it}
  - FE is biased and inconsistent if u_{it} is correlated with any of the explanatory variables in any time period
  - To see this, remember that the RHS variable in the transformed model is

$$\ddot{x}_{it} = x_{it} - T^{-1} \sum_{t=1}^{T} x_{it}$$

  which includes the x_{it} for every time period

- However, just as the FD estimator does, FE allows for arbitrary correlation between the explanatory variables and a_i

FD vs FE

- In the two-period case, FD and FE are the same
- This is not true for $T \geq 3$
- Both consistent under strict exogeneity
- Which should you use?
  - The answer is a bit tricky
  - Depends on what we expect the serial correlation in error terms to look like
- In the typical micro-data panel case, we have short T and large N
  - Typically end up preferring FE
  - Also, you throw away less information than in FD

Least Squares Dummy Variable Implementation

- Usually, we treat the fixed effects as nuisance parameters and just want to be rid of them
- Oftentimes the key issue is just about solving omitted variable bias
- Occasionally, however, we may be interested in the actual value of the fixed effect
  - E.g., models of teacher and school effectiveness
  - Or perhaps in seeing the means in particular groups
- An alternative to the within-transformation is to include a dummy variable for every value of the grouping variable
  - E.g., you may put in year dummy variables or subtract the year-specific mean from all variables
  - The estimates will be identical!
  - However, the R^2 is not directly comparable...

How to Do This in Practice?

- So, you can just include indicators for each unit i in your regression
- We typically denote this as follows:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + \alpha_i + u_{it}$$

  where α_i are the "unit-specific fixed effects"

- On Stata, use e.g. `reghdfe`

Controlling for Fixed Effects Can Make a Big Difference in Practice

Four estimates of the effects of unionization, from Mostly Harmless Econometrics

Table 5.1.1: Estimated effects of union status on log wages

Survey	Cross section estimate	Fixed effects estimate
May CPS, 1974-75	0.19	0.09
National Longitudinal Survey of Young Men, 1970-78	0.28	0.19
Michigan PSID, 1970-79	0.23	0.14
QES, 1973-77	0.14	0.16

Notes: Adapted from Freeman (1984). The table reports cross-section and panel estimates of the union relative wage effect. The estimates were calculated using the surveys listed at left. The cross-section estimates include controls for demographic and human capital variables.

Standard Errors in FE Specifications

- Thinking of inference, we face a problem that standard errors in the panel setting are often serially correlated
  - Think of this as a clustering problem
  - The problem applies more generally to panel data, not just FE specifications
  - Standard errors are clustered at the (say) state level

- Not accounting for this serial correlation makes the standard error estimates in panel regressions, such as difference-in-difference specifications, too small (usually)

- **General advice:** cluster your standard errors at the right level (state/county/village/individual/firm/...), conditional on there being enough groups

An Empirical Example: Booms, Busts and Babies (Dehejia and Lleras-Muney 2004)

- Dehejia and Lleras-Muney study the effects of the unemployment rate in different states on the birth rate and the health of children born
- Their core specification is:

$$Y_{st} = \alpha + \beta \mathit{unemployment}_{st} + \rho_s + \theta_t + \rho_s t + \epsilon_{st}$$

where Y_{st} is the outcome in state s at time t, the variable t is a year trend, and ρ_s is a vector of state FE, θ_t is a vector of year FE

Booms, Busts and Babies: Unemployment and Birth Rates

TABLE II
EFFECT OF UNEMPLOYMENT ON BIRTHRATE AND PERCENT BLACK

Dependent variable	(1) Overall birthrate	(2) White birthrate	(3) Black birthrate	(4) Overall birthrate	(5) White birthrate	(6) Black birthrate	(7) % black babies	(8) % black babies
Unemployment rate	0.000096	0.00019	−0.00047	−0.00022	−0.00019	−0.00047	−0.0018***	−0.00059***
	(0.00034)	(0.00038)	(0.00039)	(0.00023)	(0.00025)	(0.00032)	(0.00038)	(0.00028)
% effect	*0.0015*	*0.013571*	*−0.02474*	*−0.034*	*−0.00792*	*−0.01621*	*−0.0125*	*−0.0041*
State fixed effects	X	x	x	x	x	x	x	x
Year fixed effects	X	x	x	x	x	x	x	x
State-specific trend				x	x	x		x
Observations	2506	1253	1253	2506	1253	1253	1253	1253
R^2	0.51	0.58	0.55	0.74	0.77	0.79	1.00	1.00

Birthrate data are by state, year, and race. Birthrate = number of births divided by population by state and year. Percent Black babies is the ratio of Black births to total births by state and year. Births are matched to unemployment rates by state and year of conception. All regressions are weighted using the number of births in the state, year, and race as weights. Robust standard errors are in parentheses. * significant at 10 percent; ** significant at 5 percent; *** significant at 1 percent.

Booms, Busts and Babies: Children's Health Outcomes

TABLE IIIa

EFFECT OF UNEMPLOYMENT ON CHILDREN'S HEALTH OUTCOMES

Dependent variable	(1) % born below 2500 grams	(2) % born below 1500 grams	(3) % with Apgar score 5 and below	(4) Infant mortality rate	(5) Neonatal mortality rate	(6) Postneonatal mortality rate	(7) Congenital defects
All mothers							
Unemployment rate	−0.00034***	−0.00006*	−0.00003	−6.549***	−2.825	−3.726***	0.00009
With state and year fe	(0.000064)	(0.000033)	(0.000045)	(2.336)	(1.829)	(0.933)	(0.00013)
% effect of 1% Δ in u-rate	*−0.50%*	*−0.46%*	*−0.30%*	*−0.67%*	*−0.44%*	*−1.10%*	*0.69%*
Unemployment rate	−0.00018***	−0.00007*	−0.000024	−4.940*	−1.825	−3.117***	0.00011
With state and year fe, and state trends	(0.000063)	(0.00003)	(0.00005)	(2.657)	(2.039)	(1.134)	(0.00015)
% effect of 1% Δ in u-rate	*−0.26%*	*−0.54%*	*−0.24%*	*−0.51%*	*−0.29%*	*−0.92%*	*0.85%*

Data from the Natality Files are aggregated to the state, year, and race level, for states and years as listed in Appendix 1. Child mortality data are by state and year for 1979–1998. Infant mortality rates are computed as the number of infants who die within a year of birth as a fraction of live births * 1000, and likewise for neonatal mortality (the number of infants who die within 28 days) and postneonatal mortality (number of infants who die between 28 days and a year of birth). All regressions include state and year fixed effects. The unemployment rate is calculated at the state-year level and matched to the Natality Files (birth weight, Apgar score) by the year of conception of the baby and to mortality data by the year prior to child mortality. They are weighted by the number of births in the state. Robust standard errors are in parentheses. For the difference between Black and White, the *p*-value tests whether the unemployment-race coefficient is significantly different from zero in a model that is fully interacted with race. * significant at 10 percent; ** significant at 5 percent; *** significant at 1 percent.

Readings

Important

- Wooldridge (2013) Introductory Econometrics: A Modern Approach, Chapter 13 and 14 (including the appendix—but feel free to skip the parts about random effects)
- Angrist and Pischke (2009) Chapter 5 (Sections 5.1 and 5.2)

Other references mentioned in these lecture notes

- Jayachandran, S., Lleras-Muney, A., & Smith, K. V. (2010). Modern medicine and the twentieth century decline in mortality: evidence on the impact of sulfa drugs. American Economic Journal: Applied Economics, 2(2), 118-146.
- Rajeev Dehejia, Adriana Lleras-Muney, Booms, Busts, and Babies' Health, The Quarterly Journal of Economics, Volume 119, Issue 3, August 2004, Pages 1091–1130,