Intuition to Panel Data

Mary Kaltenberg
mkaltenberg@pace.edu

April 16, 2021

Data Type: What is Panel Data?

- Follows entities over time.
  - This may be states, individuals, households or countries,etc.
- Large number of observations and few time periods (T<N)
- Type of longitudinal data
- Most software require the data to be in 'long' format

nr[1] 13

	nr	year	agric	black	bus	construc	ent	exper	fin	hisp	poorhlth	hours	manuf	marr
1	13	1980	0	0	1	0	0	1	0	0	0	2672	0	
2	13	1981	0	0	0	0	0	2	0	0	0	2320	0	
3	13	1982	0	0	1	0	0	3	0	0	0	2940	0	
4	13	1983	0	0	1	0	0	4	0	0	0	2960	0	
5	13	1984	0	0	0	0	0	5	0	0	0	3071	0	
6	13	1985	0	0	1	0	0	6	0	0	0	2864	0	
7	13	1986	0	0	1	0	0	7	0	0	0	2994	0	
8	13	1987	0	0	1	0	0	8	0	0	0	2640	0	
9	17	1980	0	0	0	0	0	4	0	0	0	2484	0	
10	17	1981	0	0	0	0	0	5	0	0	0	2804	0	
11	17	1982	0	0	0	0	0	6	0	0	0	2530	0	
12	17	1983	0	0	0	0	0	7	0	0	0	2340	0	
13	17	1984	0	0	0	0	0	8	0	0	0	2486	0	
14	17	1985	0	0	0	1	0	9	0	0	0	2164	0	
15	17	1986	0	0	0	1	0	10	0	0	0	2749	0	
16	17	1987	0	0	0	1	0	11	0	0	0	2476	0	
17	18	1980	0	0	0	0	0	4	0	0	0	2332	0	
18	18	1981	0	0	0	0	0	5	0	0	0	2116	0	
19	18	1982	0	0	0	0	0	6	0	0	0	2500	1	
20	18	1983	0	0	0	0	0	7	0	0	0	2474	0	
21	18	1984	0	0	0	0	0	8	0	0	0	2362	0	
22	18	1985	0	0	0	0	0	9	0	0	0	2340	0	
23	18	1986	0	0	0	0	0	10	0	0	0	2340	0	
24	18	1987	0	0	0	0	0	11	0	0	0	2340	0	
25	45	1980	0	0	0	0	0	2	0	0	0	1864	0	
26	45	1981	0	0	0	0	0	3	0	0	0	2021	0	
27	45	1982	0	0	0	0	0	4	0	0	0	2274	1	
28	45	1983	0	0	0	0	0	5	0	0	0	2112	1	
29	45	1984	0	0	1	0	0	6	0	0	0	1920	0	
30	45	1985	0	0	0	0	0	7	0	0	0	2205	1	
31	45	1986	0	0	0	0	0	8	0	0	0	2551	0	
32	45	1987	0	0	0	1	0	9	0	0	0	2860	0	

Variables

Enter filter text here

Name	Label
nr	person identifier
year	1980 to 1987
agric	=1 if in agriculture
black	=1 if black
bus	=1 if business & re...
construc	=1 if in construction
ent	=1 if entertainment
exper	labor mkt experience
fin	=1 if finance
hisp	=1 if Hispanic
poorhlth	=1 if in poor health
hours	annual hours worked
manuf	=1 if in manufactur...
married	=1 if married

Properties

Variables	
Name	nr
Label	person identifier
Type	int
Format	%9.0g
Value Label	
Notes	
Data	
Filename	wagepan.dta
Full Path	/Users/mkaltenberg/Dropbox
Label	
Notes	
Variables	44
Observations	4,360
Size	217.15K

Flavors of Panel Data

- Micro Datasets - follows individuals, households, or firms over time (usually shorter time periods)
  - National Longitudinal Survey
  - Household Panel Survey
  - Census or Administrative Data
  - DHS Surveys
- Macro Datasets - follows countries or regions over time (usually longer time periods utilizing aggregated data)
  - Penn World Tables
  - World Development Indicators
  - Demographic Yearbook (UN)
  - Eurostat

These require different econometric techniques

- Macro Dynamics are much different than micro Dynamics
- Impact of time is much more severe in macro
- Limited number of observations in macro data

Pooled OLS

Suppose this **model**:

$$\hat{lnwage}_i = \hat{\beta}_0 + \hat{\beta}_1 Union_i + \hat{\beta}_2 Exper_i + \hat{\beta}_3 Exper_i^2 + u_i \quad (1)$$

Which we **estimate** with OLS

This averages everything together (no distinction of time periods)

*Note the **difference** between model and estimation method

Why use Panel Data Models?

We live in a **complex world** - Many **unobservable** variables can affect our outcome (Violates covariates are exogenous)

$$E(u_i|X_{1i},\ldots,X_{Ki}) = 0 \quad \forall i,t.$$

We live in a **data accessible world** - We can follow individuals over time (Violates *iid*, identically and independently distributed)

$$(X_{i1},\ldots,X_{iT},u_{i1},\ldots,u_{iT}), \quad i=1,\ldots,N \quad \text{are} \quad i.i.d.$$

Each observation is **independent**ly drawn from a random sample

We live in a **dynamic world - time** can have a big impact (Violates constant variance and autocorrelation)

$$Var(u_i|X_{1i},\ldots,X_{Ki}) = \sigma_u^2 \quad \forall t=1,\ldots,T$$

$$\mathrm{E}(u_{it},u_{is}|X_{it},X_{is},\alpha_i) = 0, \quad \forall t \neq s.$$

Panel Data - Fixed Effects

To overcome these issues, we can use a **Panel Data**

Let's start with how to deal with a complex world - **unobservable variables**

If we don't include variables that explain our dependent variable, the estimate of the coefficient of interest can suffer from *Omitted Variable Bias*.

Fixed Effects can remove biasedness caused by **time-constant** unobservable characteristics (Often why it is a popular model).

For example, IQ or work ethic can influence wages. These characteristics are very hard to empirically measure, but do not change in an individual over time.

Panel Data - First Differences

We can use time to our advantage and compare “before" and “after" when $T = 2$

We take the difference of time period 2 and time period 1

$$lwage_{i,1981} = \beta_0 + \beta_1 union_{i,1981} + \beta Z_i + u_{i,1981} \tag{2}$$

$$lwage_{i,1987} = \beta_0 + \beta_1 union_{i,1987} + \beta Z_i + u_{i,1987} \tag{3}$$

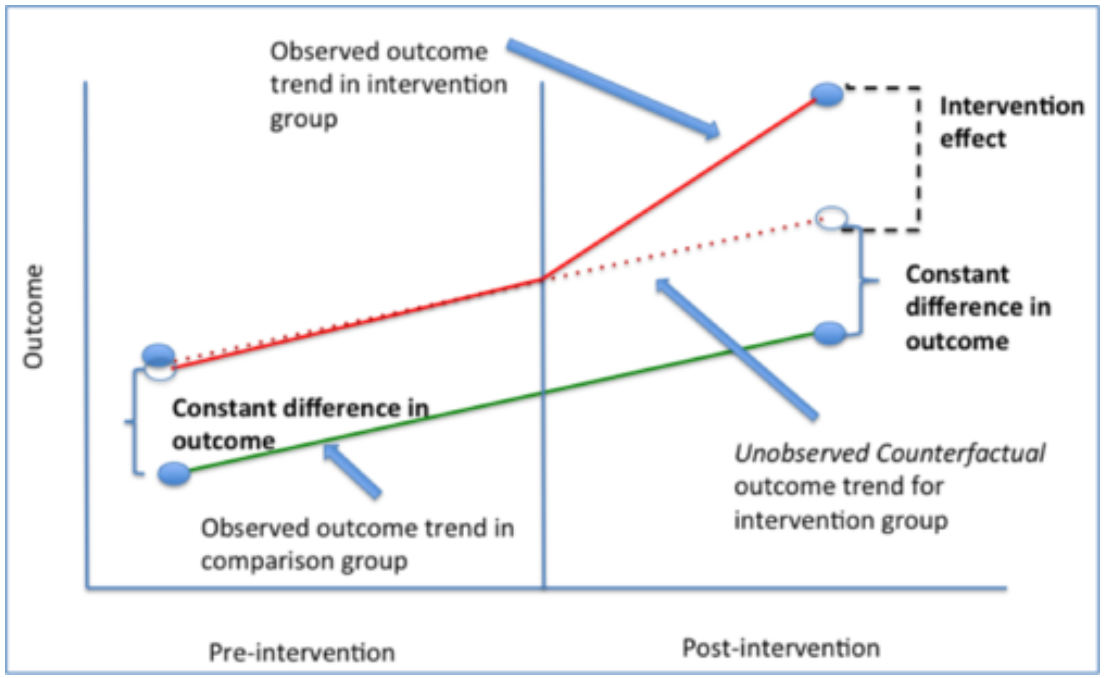
Subtracting 2 and 3 simplifies to:

$$lwage_{i,1987} - lwage_{i,1981} = \beta_1 (union_{i,1987} - union_{i,1981}) - u_{i,1987} - u_{i,1981} \tag{4}$$

This removes all of the time constant observable characteristics (Z), but we are left with the **variable of interest** (union membership) and the **error term**.

Difference-in-Differences: A Note

Difference-in-Differences (A&P 5) also makes use of time differences. The key to this method is to identify a proper **control** and **treatment** group where an **event** takes place between the two time periods. Sequence or time ordering of events can help us identify causality (but, not in all cases)



Fixed Effects

Entity Fixed Effects (one-way) model is:

$$lwage_{it} = \beta_1 union_{1,it} + \beta_2 exper_{2,it} + \beta_3 expersq_{3,it} + \alpha_i + u_{it} \qquad (5)$$

where $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. The α_i are individual level effects that are constant across time. *N.B: this model does not have a constant.*

There are n different intercepts for each entity. For example, there is an intercept for each individual in this example. Recall that these entities can be anything from countries, states, industries, households to school, etc.

Least Square Dummy Variable Model

To help understand this model, let's look at it in a different perspective. Sometimes this model is called Least Square Dummy Variable.

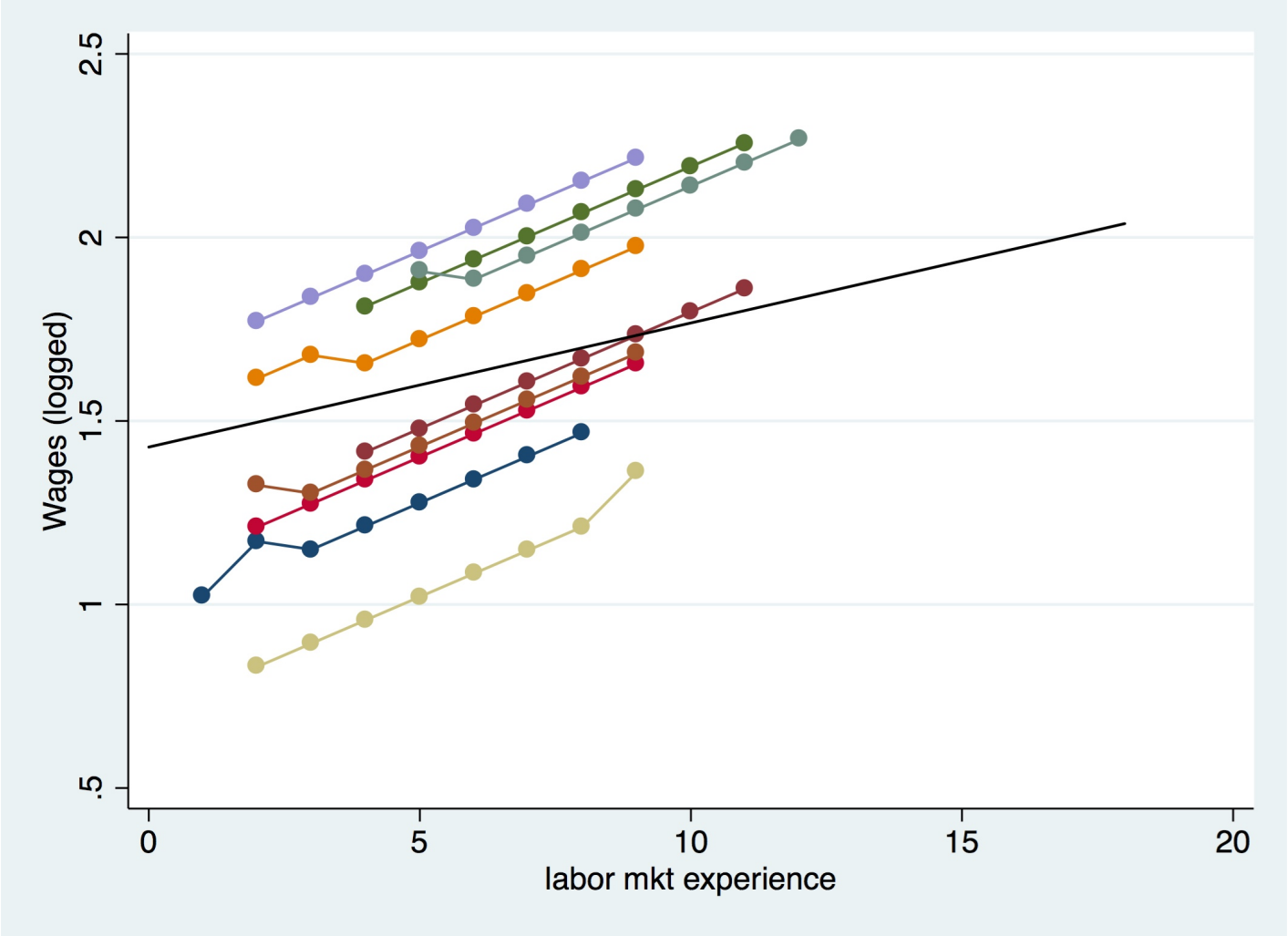
Suppose we include a binary variable for each entity (excluding one as the reference and to avoid perfect multicollinearity)
Out Model becomes:

$$lwage_{it} = \beta_0 + \beta exper_{it} + \beta expersq_{it} + \gamma_2 D2_i + \cdots + \gamma_n Dn_i + u_{it} \quad (6)$$

Model 5 and 6 are equivalent.

OLS vs. FE

Visual comparison:



Note on Estimating (FE)

Usually FE is not estimated using the LSDV model (7).

Define the within transformation as:

$$\tilde{lwage}_{it} = lwage_{it} - \bar{lwage}_i \tag{7}$$

$$\tilde{exper}_{j,it} = exper_{j,it} - \bar{exper}_{j,i} \tag{8}$$

Then we can run a transformed regression:

$$\tilde{lwage}_{it} = \beta_1 \tilde{union}_{1,it} + \beta_2 \tilde{exper}_{2,it} + u_{it} \tag{9}$$

This highlights that the estimated fixed effects α_i are unit level means (after partialling out the Xs (union and exper)).
We are differencing with respect to the mean. Let's see how Model 4, 6 and 5 are equivalent in the case where T = 2.

Fixed Effects with Time

Let's tackle another issue: we live in a dynamic world - **time** can have a big impact.

There may be **variables that change over time affecting wages for all individuals**, such as economic growth or a change in the national minimum wage. We can control for these changes over time that effect all entities by including time intercepts for each year.

Time may also affect how we **calculate our standard errors**. If our past values influence our future values, this will cause serial correlation. We can adjust our standard errors with HAC (heteroskedasticity and autocorrelation-consistent) robust standard errors.

Fixed Effects with Time

The Entity and Time Fixed Effects (two-way) model is:

$$lwage_{it} = \beta_1 union_{it} + \beta_2 exper_{it} + \beta_3 expersq_{it} + \alpha_i + \lambda_t + u_{it} \quad (10)$$

The α_i are individual level effects that are constant across time and the λ_t are time level effects that are constant across individuals.

Assume $T_i = T$ for all i, so that the panel is "balanced".

Q: How might this model look like in the LSDV model format?

Fixed Effects with Time: Autocorrelation

Autocorrelation (or serial correlation) is the correlation of a variable with itself in different time periods. Given you observe some variable (say, z_t) over time, autocorrelation is defined as

$$\text{Corr}(z_t, z_s) = \frac{\text{Cov}(z_t, z_s)}{\sqrt{\text{Var}(z_t)\text{Var}(z_s)}} = \frac{\text{Cov}(z_t, z_s)}{\text{Var}(z_t)} \quad \forall t \neq s$$

In panel data models, we usually assume

$$\text{E}(u_{it}, u_{is} | X_{it}, X_{is}, \alpha_i) = 0, \quad \forall t \neq s.$$

This assumption is testable (also in pooled OLS where there are no α_i). Violations are easily adjusted for with *HAC* robust standard errors, but – as always – we lose some efficiency with robust methods.

Exogeneity, lagged effects and feedback

In the FE model, the crucial assumption now is

$$E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0.$$

What does this mean?

- $X_{i1}, \ldots, X_{iT}$ is the entire history of X_{i1} to X_{it} and ...
- the future $(X_{i,t+1}, \ldots, X_{iT})$, so no feedback from u_{it} to future regressors $X_{i,t+1}, \ldots, X_{iT}$ is allowed.
- No additional lagged effects than those that are modeled (the model is dynamically complete).
- All Xs are strictly exogenous (conditional on the α_is).

With time series we may want to include lags of the included regressors to allow for slow adjustment (or to limit endogeneity when we remove the contemporaneous X_{it} and keep lags only).

What FE Can't Solve

- Measurement Error
- Can't measure time constant variables (gender, race, etc).
- Time-varying unobservable effects (characteristics that change over time and do not affect all entities the same)
- simultaneity/feedback loops
- Poor estimates if there is little variation
- Only looks at within-unit change; can't evaluate between-unit variation

Random Effects Model

Sometimes we are interested in time-constant characteristics (race, gender, etc.) or between-entity variation (differences between countries or regions, etc).

Random Effects models is another panel data method that can address these concerns. However, it comes with a big caveat

The assumption that a_i is uncorrelated with any of the explanatory variables. This is written as:

$$Cov(x_{itj}, a_i) = 0, t = 1, 2, \ldots, T; j = 1, 2, \ldots, k$$

Suppose the Random Effects model:

$$lwage_{it} = \beta_0 + \beta_1 union_{it} + \beta_2 exper_{it} + \beta_4 black_i + \beta_4 hisp_i + a_{it} + u_{it} \quad (11)$$

In this case, a_i are the intercepts for individual characteristics, black and hispanic. We assume these variables are not correlated with experience or union membership.

The Random Effects model partially demeans each variable

The a_i are treated as random variables, rather than fixed constants.
Let's see how it works
N.B. Random Effects is usually estimated using FGLS (Feasible Generalized Least Squares).
Often, you will see

$$a_{it} + u_{it} = v_{it}$$

which is called a composite error term

Random Effects vs. Fixed Effects

This assumption can often be violated and it may be hard for the researcher to know. Luckily, there is a statistical test to help guide us, called the **Hausman test**

- Compares the estimated coefficient of RE and FE
- Tests if the differences between these estimates are statistically significant
- Rejecting the null indicates that individual specific effects are uncorrelated with regressors (ie if p-value is low, then use FE model)

Note on Endogeneity

We assume that the independent variables and the error term are uncorrelated:
$E[x_i|u_i] = 0$ (zero conditional mean assumption)

If the independent variables and error term are correlated, this is a problem. Much of econometrics is constantly trying to solve this issue.

Note on Endogeneity

Endogeneity usually arises through:

1. Omitted variables
   - Variables that we would like to control for, but cannot because of data availability
   - Often arises due to self-selection; e.g. individualâĂŹs decision on years of schooling is likely to be correlated with unobserved ability
   - This is the main problem associated with estimating treatment effects
2. Measurement error
   - Where we only observe an imperfect measure of our variable of interest. The measurement error forms part of the error term
3. Simultaneity
   - Arises when at least one explanatory variable is determined simultaneously with the dependent variable,
   - In such cases there is generally a correlation between the simultaneously determined explanatory variable and the error term