4. Panel data

LPO 8852: Regression II

Sean P. Corcoran

Difference-in-differences recap

Difference-in-differences (DD) often relies on *panel data*, with repeat observations of two or more groups (i) over time (t).

$$y_{it} = \beta_0 + \beta_1 treat_i + \beta_2 post_t + \beta_3(treat_i \times post_t) + \gamma X_{it} + u_{it}$$

$treat_i$ is a fixed effect for the treated group, $post_t$ is a time effect equal to one in the post period(s). Regression DD often includes year effects as well.

Difference-in-differences recap

What identification problem does DD solve? Treated observations may differ systematically from untreated observations (β_1)—differencing over time "nets out" these fixed differences and focuses on changes over time. β_2 is intended to capture the change over time that would have occurred for treated observations had they not been treated.

OVB remains if there are omitted variables correlated with *treat* × *post* and the outcome y. For example: unobserved factors that change differentially for treated observations (implying non-parallel trends).

Difference-in-differences recap

DD was our first attempt to address selection on *unobservables*. Treatment need not be randomly assigned, and treated and untreated units *can* differ systematically prior to treatment (this is captured by β_1, or the group specific coefficients in the generalized DD).

As long as these unobserved differences do not change over time, DD can eliminate the unobserved selection bias.

Panel data

Panel, longitudinal, or *"cross-sectional time series"* data consist of observations on cross-sectional units (e.g., students, schools, hospitals, neighborhoods, counties, states) at multiple points in time.

- N cross-sectional (panel) units and T time periods ($T \geq 2$)
- A *balanced panel* has exactly $N \times T$ observations (T time observations for all N panel units)
- An *unbalanced panel* has T_i observations for panel unit i, where T_i is not the same for all i

Differs from a *pooled cross-section*, although panel methods can be used with this type of data (e.g., Kearney & Levine (2019) example)

Panel data - long

Panel data in *long* format, N students in $T = 4$ years:

studentID	year	readscore	mathscore	incomecat	...
1	1999	75	82	3	
1	2000	78	84	4	
1	2001	80	90	4	
1	2002	78	91	3	
2	1999	91	92	2	
2	2000	94	92	2	
2	2001	80	85	2	
2	2002	87	83	2	
3	1999	62	50	5	
3	2000	70	47	5	
3	2001	75	55	4	
3	2002	73	60	5	

Panel data - wide

Panel data in *wide* format, N students in $T = 4$ years:

studentID	read99	math99	inc99	read00	math00	inc00	read01	...
1	75	82	3	78	84	4	80	
2	91	92	2	94	92	2	80	
3	62	50	5	70	47	5	75	
4	...	...	...					

Panel data - reshape long

Moving between *long* and *wide* format in Stata with `reshape`, beginning with *wide* data

- `i()` contains the time invariant variables (e.g., ID, gender)
- `j()` specifies the time variable to be created (e.g., year)
- The list of time varying variables are "stubs" that end in the *j* suffix

```
reshape long stubnames, i(varlist) j(varname)
```

- If `j()` consists of *string* rather than *numeric* values, use the `string` option
- Example time-varying variable names: *score98, score99, score00* (Stata may have problems with 00 as a `j()` value if `string` option is not used).

Panel data - reshape wide

Moving between *long* and *wide* format in Stata with `reshape`, beginning with *long* data

- `i()` contains the time invariant variables (e.g., ID, gender)
- `j()` specifies the time variable (e.g., year)
- The list of time varying variables are "stubs" that *will* end in the *j* suffix, once converted to wide

```
reshape wide stubnames, i(varlist) j(varname)
```

- After reshaping, Stata allows you to revert back easily without losing information. E.g., after the above command just type `reshape long`
- Most panel regression commands expect the data to be in *long* format.

In-class example 1

Illustration of reshape commands using *Census_states_1970_2000* data:

- Cross sectional unit: state
- Time variable: year (decennial Census years)
- Time-varying variables: median household income, unemployment rate

Stata panel commands

Stata has many useful `xt` commands for working with panel data. Typically these require that you first declare the data to be a panel using `xtset`:

- `xtset` *`panelvar timevar`*
- The *panelvar* must be numeric. If it is not, you can use encode: **`encode`** *`panelvar`*, `gen(`*`panelvar2`*`)`
- It is possible to tell Stata in the `xtset` options what units of time the data represent—e.g., years, quarters, minutes (useful for some purposes)
- `xtset` alone will report back the panel settings

Stata panel commands

Other useful Stata panel data commands for description:

- `xtdescribe`—to see patterns of participation/data availability
- `xtsum`—for descriptive statistics that show between- and within-unit variation
- `xttab`—for one-way tabulations with separate counts within and between units
- `xttrans`—for transition probabilities (movement between categories of a categorical variable)
- `xtline` and `xtline, overlay`—for separate line graphs by panel unit (see in-class example)

xtsum

Decomposition of variation in xtsum:

$$s_w^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x}_i)^2$$

$$s_b^2 = \frac{1}{N-1}\sum_i (\bar{x}_i - \bar{x})^2$$

$$s_o^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x})^2$$

Note $\bar{x}$ is the *grand mean* of x. Can also write:

$$s_w^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x}_i + \bar{x})^2$$

because adding a constant ($\bar{x}$) will not affect s_w^2

xtsum

xtsum also shows the min and max of:

- x_{it}: overall
- $\bar{x}_i$: between
- $(x_{it} - \bar{x}_i + \bar{x})$: within

Note: on xtsum, see also https://www.stata.com/support/faqs/statistics/decomposed-variances-in-xtsum/

- Between and within variation do not sum to overall
- Variance estimates are bias-corrected, multiplied by $n/(n-1)$
- With unbalanced panels, s^b is calculated using mean of panel means, not $\bar{x}$ (may not be the same)

In-class example 2

Illustration of `xt` commands using *State_school_finance_panel* data:

- Cross sectional unit: state
- Time variable: year (annual 1990-2010)
- Time-varying variables: various school finance measures

Panel data - advantages

Why use panel data?

- Can help us answer questions not possible with a cross-section or time-series approach
- Can generate *measures* not possible with cross-sectional or time series data (e.g., growth, work spells)
  - If 50% of women are working in year *t*, does this reflect 50% of women working at any given point, or 50% of women who work all the time?
- Allows us to address selection bias due to unobserved heterogeneity that is fixed over time (“fixed effects”)

Selection bias revisited

Lecture 1: interpretation of regression coefficients as causal is often complicated by selection bias. Example:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

with $E(u_i|x_i) \neq 0$ because we believe potential outcomes are not independent of x. We can attempt to mitigate selection bias through the inclusion of additional covariates or via matching, but this only solves the problem if conditioning on these observables (or the propensity score) eliminates OVB.

In practice we are often more concerned about selection on *unobservables*.

Unobserved heterogeneity

Suppose there are unobserved, fixed differences across units (c_i) that affect the outcome and are (potentially) correlated with the explanatory variable of interest (x_i):

$$y_i = \beta_0 + \beta_1 x_i + c_i + u_i$$

c_i could represent the effects of ability, health, motivation, intelligence, parental resources, managerial quality, organizational culture, state/local policies or regulations, etc.

First difference model

Suppose we have two time periods (T=2) for each cross-sectional unit *i*, and assume the linear model above applies in both periods:

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + c_i + u_{i2}$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + c_i + u_{i1}$$

Now subtract period 1 from period 2 for the "first difference":

$$\Delta y_i = \beta_1 \Delta x_i + \Delta u_i$$
$$y_i^* = \beta_1 x_i^* + u_i^*$$

Because c_i is time-invariant, it differences out of the model. Notice the constant β_0 also differences out.

First difference model

The first difference model can be estimated using OLS, as long as the usual OLS assumptions apply to it:

- The new error term $u_i^* = \Delta u_i$ is uncorrelated with the new explanatory variable, $x_i^* = \Delta x_i$.
- This requires that we have no cross-period correlations between u and x: this is called **strict exogeneity**.
- The x_i must vary over time for at least some *i*, else they difference out.

In-class example 3

Example using panel of Texas elementary schools:

- use *Texas_elementary_panel_2004_2007.dta*
- xtset *campus year*
- xtdescribe
- rename *ca311tar avgpassing*
- egen *avgclass* = rowmean(*cpctg01a-cpctgmea*)
- reg *avgpassing avgclass* if *year*==2007 (cross-sectional regression for 2007)

Note: *avgclass* is the mean class size across grades, and *avgpassing* is the school average passing rate across grades and subjects.

In-class example 3

Having declared the dataset as a panel, Stata recognizes the **d.** prefix as a "difference operator":

- reg *d.avgpassing d.avgclass* if *year*==2007, noconstant

- This is the first difference regression, using 2007 only (and its lag in the calculation of *d.avgpassing* and *d.avgclass*)

- d. can be used after xtset or tsset (time series set)

- Note suppression of the constant. In theory the constant term differences out. In practice can still estimate with a constant, which allows for a year-to-year time trend.

In-class example 3

A few things to note in example 3:

- Change in coefficient on class size: does it make sense?
- Change in sample size (re: unbalanced panel due to missing values)

A few things to think about:

- Is strict exogeneity likely to hold in this circumstance?
- Where is the identifying variation coming from?
- How much variation is there in the *change* in passing rates (Δy) and class size (Δx)?
- Do outliers dominate the variation in changes?

```
gen davgpassing = d.avgpassing
/* create variable containing FD that can be described */
```

In-class example 3

The first difference model is easily generalizable to multiple years ($T > 2$).

- Each year of data is differenced with the prior year
- 1st period is sacrificed
- Must continue to think about OLS assumptions, e.g. strict exogeneity

```
reg d.avgpassing d.avgclass, noconstant
table year if e(sample)
* note 1st year of data is not used
```

Fixed effects model

Alternatively, in the *(one-way) fixed effects* model, we treat the c_i as parameters to be estimated:

$$y_{it} = \beta_0 + \beta_1 x_{it} + c_i + u_{it}$$

Effectively we are allowing for a *unique intercept* for every cross-sectional unit *i*. This is feasible to estimate since each *i* is observed multiple times.

"Least squares dummy variables" approach

Now we are estimating the intercept β_0, slope β_1, and $(N-1)$ intercepts, the "fixed effects." This can be done by including $N-1$ dummy variables in the regression, sometimes called the "least squares dummy variable" (LSDV) model:

- `reg` *avgpassing avgclass* `i.`*campus*
- For this example limit to *year*>=2006 and *houston*==1 so that the number of schools is manageable.
- `areg` is equivalent but suppresses the $(N-1)$ coefficients
- `areg` *avgpassing avgclass*, `absorb(`*campus*`)`

Note omission of first cross sectional unit with `i.campus`. You can control which unit is omitted if desired. See in-class example for interpretation.

“Least squares dummy variables” approach

There are a number of reasons why you might not want to do it this way:

- Could be time-consuming and harder on memory with large datasets (re: you are creating dummy variables for each unique i)
- Soaks up degrees of freedom; may result in the number of regressors exceeding the number of observations
- Often we are not interested in the estimates of the fixed effects themselves, so there is no need to see/report them.
- Exception: recent “school effects” and “teacher effects” studies work explicitly with fixed effects estimates ($\hat{c}_i$)—more on this later

Within transformation

Suppose that panel data are available with multiple observations per i and the model is:

$$y_{it} = \beta_0 + \beta_1 x_{it} + c_i + u_{it} \quad t = 1, ..., T \quad \forall i$$

Within each panel unit i, take the average over t on both sides and subtract the average from each it observation:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + c_i + \bar{u}_i$$

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

This is called “de-meaning” or the “within” transformation (sometimes denoted $\ddot{y}_i$). Notice that the intercept β_0 and the c_i “difference out.” c_i differences out only if it is *time invariant*.

Within transformation

Under certain assumptions, an OLS regression of the de-meaned y on the de-meaned x will yield unbiased and consistent estimates of β_1.

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

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

This is also known as the fixed effects or "within" regression, and extends to more than one explanatory variable $(x_1, ..., x_k)$.

Explanatory variables x_j that are time *invariant* fall out of the model. (They all equal their within-group mean, so the within-transformation equals zero). Examples: gender, race or ethnicity, birthplace ...

Within transformation

The fixed effects or "within" regression model can be estimated using OLS using `xtreg`:

- `xtreg` *avgpassing avgclass*`, fe`
- Note `xtset` must have been declared, or specify the cross-sectional unit in the options, e.g. `i(`*campus*`)`
- While the fixed effects are not estimated directly, can "back out" a prediction: $\hat{c}_i = \bar{y}_i - \bar{x}_i\hat{\beta}_1$
- `predict` *schlfe*`, u`

Note the estimated fixed effects from `xtreg` are not the same as the dummy coefficients from the LSDV model.

Within transformation

The command `xtdata, fe` can be used to transform your data using the within-transformation. However, this is rarely done (in my experience) since it transforms the variables in your dataset. `xtreg` will do the transformation on the fly without affecting your dataset.

Fixed effects model

Compare `xtreg`, `areg` and first difference when $T = 2$

- `xtreg` *avgpassing avgclass3*, `fe`
- `areg` *avgpassing avgclass3*, `absorb(`*campus*`)`
- `reg d.`*avgpassing* `d.`*avgclass3*, `noconstant`

Fixed effects models

A few notes about `xtreg, fe`

- FE is more efficient (smaller standard errors) than first differencing if the error terms are serially uncorrelated and $T > 2$
- Assumes no correlation in u across units of panel i (some tests for this using user-written `xtscd, xttest3`)
- The estimates of the fixed effects themselves (c_i) are unbiased but inconsistent in large samples. (Why? As the number of panel units grows ($N \to \infty$) the number of parameters to estimate also grows).
- `xtreg` has not historically allowed `svy` specification (for complex sampling designs) but can use `pweights` and `cluster()` option. See also the `mixed` (or `xtmixed`) command for an alternative.

Fixed effects model

Stata actually fits the following model with `xtreg`:

$$(y_{it} - \bar{y}_i + \bar{y}) = \beta_0 + \beta_1(x_{it} - \bar{x}_i + \bar{x}) + (u_{it} - \bar{u}_i + \bar{u})$$

Where the values with a bar but no subscript are the grand means. This includes an intercept which is the average of the fixed effects (c_i).

Fixed effects model

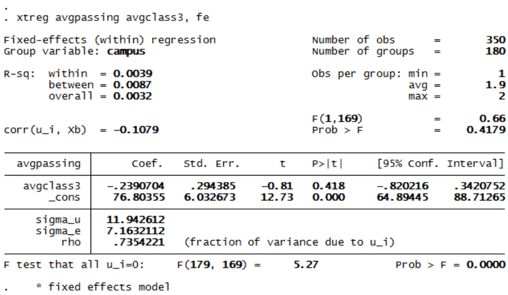
Fixed effects considerations:

- Where is the identification coming from?
- How much variation is there within panel units? When small, one risks imprecise estimates
- For stats on within- and between- school variation can use `xtsum` (described earlier), `xttrans` for categorical variables

`xtsum` *`avgpassing avgclass`*

Fixed effects model

Other useful output from `xtreg`:



```
.
. xtreg avgpassing avgclass3, fe

Fixed-effects (within) regression               Number of obs      =       350
Group variable: campus                          Number of groups   =       180

R-sq:  within  = 0.0039                         Obs per group: min =         1
       between = 0.0087                                        avg =       1.9
       overall = 0.0032                                        max =         2

                                                F(1,169)           =      0.66
corr(u_i, Xb)  = -0.1079                        Prob > F           =    0.4179

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   avgclass3 |  -.2390704    .294385    -0.81   0.418     -.820216    .3420752
       _cons |   76.80355   6.032673    12.73   0.000     64.89445    88.71265
-------------+----------------------------------------------------------------
     sigma_u |  11.942612
     sigma_e |  7.1632112
         rho |   .7354221   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(179, 169) =      5.27             Prob > F = 0.0000

.     * fixed effects model
```

Fixed effects model

Other useful output from `xtreg`:

- F-test for joint significance of fixed effects (null hypothesis H_0 is that all fixed effects are zero). If rejected, fixed effects model is a reasonable assumption and regular OLS may provide inconsistent estimates. In practice, rarely rejected.
- R^2 *within*: variance "explained" by within-group deviations from group means
- R^2 *between*: variance in group means $\bar{y}_i$ "explained" by the group mean x's: $\bar{x}_i$
- *sigma_u* estimate of the standard deviation in fixed effects (c_i)

Fixed effects model: assumptions for inference

- **FE.1:** linear model $y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + c_i + u_{it}$
- **FE.2:** cross-sectional units are a random sample
- **FE.3:** x_{it} varies over time for some i, no perfect collinearity
- **FE.4:** $\forall t$, $E(u_{it}|X_i, c_i) = 0$ or the expected value of u given x in *all* time periods is zero (strict exogeneity)
- **FE.5:** $Var(u_{it}|X_i, c_i) = Var(u_{it}) = \sigma_u^2$ - homoskedasticity
- **FE.6:** for $t \neq s$ errors are uncorrelated: $Cov(u_{it}, u_{is}|x_i, c_i) = 0$. No serial correlation.

Under FE.1-FE.4, fixed effects model (and first difference model) is unbiased. Adding FE.5-FE.6, fixed effects model is BLUE. If FE.6 holds, fixed effects is more efficient than the first difference model. Can relax homoskedasticity assumption and calculate `robust` standard errors.

Fixed effects model: assumptions for inference

Note: the econometric theory described here is for “short” panels, with N large relative to T. If the opposite is true in your context, use FE model with caution (see Wooldridge chapter 14, Cameron & Trivedi).

Two-way fixed effects model

The two-way fixed effects model adds another dimension of fixed effects (often time periods). There is no explicit command for two-way models, rather can just include time dummies. Alternatively, `reghdfe`

```
xtreg avgpassing avgclass i.year, fe
* the i.year syntax introduces (T-1) time effects
test _Iyear_2006 _Iyear_2007
* joint test that time effects = 0
reghdfe avgpassing avgclass, absorb(campus year)
```

As with one-way fixed effects model, requires variation across units within time periods t.

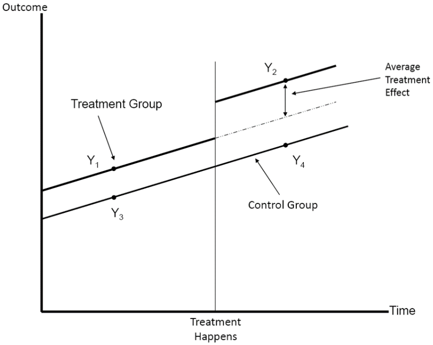
Two-way fixed effects model

The generalized difference-in-differences model is a two-way fixed effects model:

$$y_{it} = \beta_0 + \beta_1(treat_i \times post_t) + \alpha_t + \gamma_i + \delta X_{it} + u_{it}$$

There are cross-sectional unit fixed effects (γ_i) which represent separate intercepts for each unit and time effects (α_t) which represent common variation over time within group.

Two-way fixed effects model



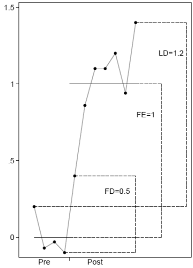
Comparison of models

It is important to be attentive to where the variation in each type of FE model is coming from:

- Fixed effects ("within") model: uses deviations from group (i) means, e.g., mean "pre" vs. mean "post"
- First differences model: uses variation in successive time periods, e.g., just prior to and just after a "treatment" (a change in x)
- **Long differences** is like first differences, but there is a long time span between observations. Here outcomes may be compared well before and well after a "treatment".

To evaluate these in your situation, need some idea of the speed in which x affects y

Comparison of models



Source: Nichols (2007). Figure shows one panel (i)'s contribution to the estimated effect of a treatment that $= 1$ in post period ($t > 4$). Notice the different treatment effects depending on FE, FD, or LD.

Fixed effects models in other applications

Fixed effects models are not exclusively used with panel data in which cross-sectional units *i* are observed in multiple time periods. They are also used with grouped or clustered data. For example:

- Family fixed effects, where the family is the cross-sectional unit and siblings are the group members (akin to the time dimension)
- School fixed effects with student-level data, where each school has its own intercept

Fixed effects models in other applications

The researcher needs to provide a convincing rationale for why the unobserved variable should be considered fixed across multiple observations (e.g., siblings, years)

- Why did a mother's employment status change between siblings?
- Why did only 1 of 2 siblings participate in Head Start?
- Why did a student switch from a traditional school to a charter school?
- Why did an elementary school receive a new principal?

Standard errors

The assumption that errors u_{it} are i.i.d. is not often satisfied in panels. With repeat observations on the same cross-sectional unit, it is likely that errors are correlated across observations for the same i.

- If y is over-predicted in one period for a given i, it is likely to be over-predicted in the next period.
- For "short" panels (large N, small T), can use cluster-robust standard errors
- The "cluster" is typically the cross-sectional unit, although when the regressor of interest is aggregated at a higher level (e.g., state), can cluster at that level. Theory requires large N and that higher levels nest the cross-sectional units.
- `vce(robust)` or `robust` in `xtreg` assumes data are clustered
- Cluster-robust standard errors from `areg` are different from those using `xtreg, fe`. It is recommended that you use `xtreg`.

Other topics, to be added

- The "between" model `xtreg, be`
- The random effects model `xtreg, re`
- The correlated random effects model
- Mixed linear model (random slopes)