AFRE 835: Introductory Econometrics

Chapter 13: Pooling Cross Sections across Time: Simple Panel Data Methods

Spring 2017

Introduction

- Up until now, we have focused on datasets that are either cross-sectional or time series.
- The next two chapters focus on datasets that vary across agents (individuals, firms, etc.) and over time.
    1. The first of these is the **independently pooled cross-section**, which consists of a sequence of cross sections sampled independently over time.
       ...In this case, the same individual is unlikely to appear in multiple time periods.
    2. The second is the **panel data** set, in which one explicitly tries to track the same set of individuals over time.
- There are variants of these two data types; e.g., a rolling panel in which individuals are followed over a time period, but replaced by a new set of individuals over time.

Outline

Pooling Independent Cross Sections Over Time

- There are many examples of pooled cross-sections.
- Research firms and government agencies will often repeat a survey at regular intervals, drawing a new set of units (individuals, firms, etc.) each time.
  - The *Current Population Survey (CPS)* is a survey of 60,000 households (has a monthly panel aspect, but households selected annualy).
  - The *American Time Use Survey (ATUS)* (2003-15) draws on a subsample of the CPS sample.
  - The *British Social Attitudes Survey* of 3,300 respondents.
- Because the cross-sections are drawn at different points in time, they will typically not be identically distributed.
- In fact, one is often interested in whether the relationships between the dependent variable of interest and the regressors are stable over time.
- At a minimum, it is prudent to allow the intercept to shift over time, as in Wooldridge's Example 13.1 examining women's fertility over time.

TABLE 13.1 Determinants of Women's Fertility

Dependent Variable: *kids*

Independent Variables	Coefficients	Standard Errors
educ	−.128	.018
age	.532	.138
age^2	−.0058	.0016
black	1.076	.174
east	.217	.133
northcen	.363	.121
west	.198	.167
farm	−.053	.147
othrural	−.163	.175
town	.084	.124
smcity	.212	.160
y74	.268	.173
y76	−.097	.179
y78	−.069	.182
y80	−.071	.183
y82	−.522	.172
y84	−.545	.175
constant	−7.742	3.052

$n = 1{,}129$
$R^2 = .1295$
$\bar{R}^2 = .1162$

Notice what happens to partial effect of education if we drop the time dummies.

```
. reg kids educ age agesq black east northcen west farm othrural town smcity

      Source |       SS           df       MS      Number of obs   =     1,129
-------------+----------------------------------   F(11, 1117)     =     11.52
       Model |  314.471892        11  28.5883539   Prob > F        =    0.0000
    Residual |  2771.03741     1,117   2.4807855   R-squared       =    0.1019
-------------+----------------------------------   Adj R-squared   =    0.0931
       Total |   3085.5093     1,128  2.73538059   Root MSE        =    1.5751

------------------------------------------------------------------------------
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1428788    .018351    -7.79   0.000    -.1788851   -.1068725
         age |   .5624223   .1396257     4.03   0.000     .2884641    .8363804
       agesq |  -.0060917   .0015793    -3.86   0.000    -.0091903    -.002993
       black |    .977559    .173188     5.64   0.000     .6377485     1.31737
        east |   .2362931   .1340365     1.76   0.078    -.0266987    .4992849
    northcen |   .3847487   .1222117     3.15   0.002     .1449583    .6245391
        west |   .2447027   .1686052     1.45   0.147    -.0861158    .5755212
        farm |   -.054186   .1486156    -0.36   0.715    -.3457832    .2374112
    othrural |  -.1670751   .1773583    -0.94   0.346    -.5150681    .1809178
        town |   .0842369   .1257038     0.67   0.503    -.1624053    .3308792
      smcity |   .1830768   .1620166     1.13   0.259    -.1348143     .500968
       _cons |  -8.487543   3.068381    -2.77   0.006    -14.50798   -2.467104
------------------------------------------------------------------------------
```

Allowing for Changes in Marginal Effects

- One might want to allow for changes in the marginal effect of a variable over time.
- For example, in examining the returns to educations, one might want to allow for
  - changes in the returns to education (reflecting changes in the educational demands of jobs)
  - changes in gender bias.
- Wooldridge considers the following model

$$\begin{aligned} log(wages) &= \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper \\ &+ \delta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u \\ &= (\beta_0 + \delta_0 y85) + (\beta_1 + \delta_1 y85) educ + \beta_2 exper \\ &+ \delta_3 exper^2 + \beta_4 union + (\beta_5 + \delta_5 y85) female + u \end{aligned} \tag{1}$$

Returns to Education

	Base	Base
educ	0.088	0.075
	(0.006)**	(0.007)**
exper	0.032	0.030
	(0.004)**	(0.004)**
expersq	-0.000	-0.000
	(0.000)**	(0.000)**
union	0.143	0.202
	(0.033)**	(0.030)**
female	-0.251	-0.317
	(0.028)**	(0.037)**
y85		0.118
		(0.124)
y85educ		0.018
		(0.009)*
y85fem		0.085
		(0.051)
_cons	0.443	0.459
	(0.083)**	(0.093)**
R2	0.30	0.43
N	1,084	1,084

* p<0.05; ** p<0.01

F-test restricting the two slope changes to zero has a p-value of 0.0326.

The Chow Test

- Earlier (ch. 7), the Chow test was introduced to test for differences between groups; e.g.,
  - Males versus females in earnings;
  - Racial groups in loan application success;
  - Regional differences in housing prices.
- The same test can be used to examine changes in the population regression function over time.
- The basic Chow test in this case takes the form

$$F = \frac{(SSR_r - SSR_{ur})/[(T-1)(k+1)]}{SSR_{ur}/[n - T(k+1)]} \tag{2}$$

  where $SSR_{ur} = \sum_{t=1}^{n} SSR_t$, with SSR_t denoting the SSR for the model using data only from time period t.
- This version of the Chow test allows all of the parameters (intercepts and slopes) to vary by time period.

Fertility Regressions

	72	74	76	78	80	82	84	Const
educ	-0.071	-0.083	-0.117	-0.177	-0.127	-0.130	-0.212	-0.143
	(0.065)	(0.049)	(0.050)*	(0.055)**	(0.045)**	(0.045)**	(0.043)**	(0.018)**
age	0.531	0.139	0.214	0.265	1.145	0.475	0.501	0.562
	(0.465)	(0.347)	(0.400)	(0.397)	(0.385)**	(0.338)	(0.329)	(0.140)**
agesq	-0.007	-0.002	-0.002	-0.003	-0.012	-0.004	-0.005	-0.006
	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)**	(0.004)	(0.004)	(0.002)**
black	0.861	1.158	2.416	0.410	1.758	1.175	0.343	0.978
	(0.574)	(0.505)*	(0.656)**	(0.628)	(0.490)**	(0.294)**	(0.428)	(0.173)**
east	0.677	-0.106	0.188	0.703	0.739	-0.041	-0.340	0.236
	(0.412)	(0.343)	(0.370)	(0.404)	(0.411)	(0.304)	(0.304)	(0.134)
northcen	0.608	0.403	0.213	0.327	0.636	0.403	0.229	0.385
	(0.418)	(0.298)	(0.353)	(0.362)	(0.316)*	(0.282)	(0.274)	(0.122)**
west	0.571	0.383	0.541	-0.227	0.268	0.288	-0.197	0.245
	(0.509)	(0.415)	(0.508)	(0.510)	(0.392)	(0.452)	(0.404)	(0.169)
farm	0.264	-0.294	-0.109	0.286	-0.246	-0.140	-0.509	-0.054
	(0.469)	(0.362)	(0.427)	(0.460)	(0.376)	(0.356)	(0.350)	(0.149)
othrural	0.173	-0.636	-0.756	-0.030	0.318	-0.032	-0.678	-0.167
	(0.573)	(0.421)	(0.514)	(0.498)	(0.457)	(0.403)	(0.466)	(0.177)
town	0.305	-0.147	-0.495	0.141	0.306	0.436	-0.121	0.084
	(0.370)	(0.305)	(0.364)	(0.378)	(0.344)	(0.294)	(0.290)	(0.126)
smcity	1.001	-0.425	0.132	0.382	-0.055	0.019	0.315	0.183
	(0.496)*	(0.428)	(0.541)	(0.455)	(0.411)	(0.378)	(0.343)	(0.162)
_cons	-7.343	1.875	-0.558	-1.728	-22.206	-8.646	-6.152	-8.488
	(10.352)	(7.634)	(8.728)	(8.701)	(8.427)**	(7.426)	(7.198)	(3.068)**
R2	0.10	0.08	0.15	0.11	0.23	0.26	0.21	0.10
N	156	173	152	143	142	186	177	1,129

* p<0.05; ** p<0.01

Chow Test (cont'd)

- The Chow Test test in this case is clearly rejected, with $F_{72,1045} = 1.799$, which has a p-value of 0.0002.
- Typically, one wants to at least retain time varying intercepts.
- Also, the test as run is not robust to heteroskedasticity.
- An alternative approach is to test the fully unconstrained model against one that constrains the slopes to be constant over time.
- The generic unconstrained model is given by:

$$y_t = (\beta_0 + \sum_{s=2}^{n} \delta_{0s}) + (\beta_1 x_{t1} + \sum_{s=2}^{n} \delta_{1s} x_{t1} \cdot D_{ts}) + \cdot + (\beta_k x_{tk} + \sum_{s=2}^{n} \delta_{ks} x_{tk} \cdot D_{ts})$$

- The constrained model takes the form

$$y_t = (\beta_0 + \sum_{s=2}^{n} \delta_{0s}) + \beta_1 x_{t1} + \cdots + \beta_k x_{tk}$$

- Heteroskedastic robust Wald statistic can be used to test the restriction.

Modified Test for the Fertility Case

- These models can be readily run in Stata using *factor* variables.
- For the unconstrained model:

```
********************************************************************************
*                                                                              *
*       Unconstrained model                                                    *
*                                                                              *
*******************************************************************************;
reg     kids educ age agesq black east northcen west farm othrural town smcity
            i.year#c.educ i.year#c.age i.year#c.agesq i.year#c.black
            i.year#c.east i.year#c.northcen i.year#c.west i.year#c.farm
            i.year#c.othrural i.year#c.town i.year#c.smcity i.year;
********************************************************************************
*                                                                              *
*       Constrained model with only time dummies                              *
*                                                                              *
*******************************************************************************;
reg     kids educ age agesq black east northcen west farm othrural town smcity
            i.year;
```

- The corresponding test statistic this case is $F_{60,1045} = 1.485$, which has a p-value of 0.003 (still rejecting the slope restrictions at any reasonable level).

Policy Analysis with Pooled Cross Sections

- Pooled cross-sections can be helpful in program/policy evaluation.
- Consider the implementation of a program at time t_1, such as mandatory pre-kindergarten throughout the state of Georgia.
- Suppose that the outcome of interest is first grade math test scores.
- One approach to estimating the impact of the program would be to use pooled cross-section data on test scores for first-grader in Georgia before ($t = t_0$) and after ($t = t_1$) the program, estimating the model

$$math_{it} = \beta_0 + \delta_1 D_{1t} + u_{it} \quad t = t_0, t_1, i \in \mathcal{G} \tag{3}$$

  where $D_{1t} = 1$ if $t = t_1$ (=0 otherwise) and $\mathcal{G}$ denotes Georgia students.
- The OLS estimators in this case are $\hat{\beta}_0 = \overline{math}_{G0} = \frac{1}{n_G}\sum_{i\in\mathcal{G}} math_{i,t_0}$ and $\hat{\delta}_1 = \overline{math}_{G1} - \overline{math}_{G0}$ where $\overline{math}_{G1} = \frac{1}{n_G}\sum_{i\in\mathcal{G}} math_{i,t_1}$.
- However, this inter-temporal difference in test scores $\hat{\delta}_1$ is potentially confounded with other changes.

Policy Analysis with Pooled Cross Sections (cont'd)

- An alternative would be to use cross-sectional data ($t = t_1$) on test scores in Georgia and South Carolina (where pre-K is not mandatory).
- In this case, we might estimate the model:

$$math_{it} = \beta_1 + \delta_1 G_i + u_{it} \quad t = t_1, i \in \mathcal{G}, \mathcal{M} \tag{4}$$

  where $G_i = 1$ if $i \in \mathcal{G}$ and $\mathcal{M}$ denotes Mississippi students.
- The OLS estimators in this case are

$$\hat{\beta}_1 = \overline{math}_{M1} = \frac{1}{n_M}\sum_{i\in\mathcal{M}} math_{i,t_1} \tag{5}$$

  and $\hat{\delta}_1 = \overline{math}_{G1} - \overline{math}_{M1}$ where $\overline{math}_{G1} = \frac{1}{n_G}\sum_{i\in\mathcal{G}} math_{i,t_1}$.
- This cross-sectional difference in test scores $\hat{\delta}_1$ is potentially confounded with other differences between Georgia and Mississippi students.

Differences-in-Differences (DID)

- The **Differences-in-Differences (DID) estimator** combines these data types, pooling cross-sections before and after the policy change in both the *treatment* area (Georgia) and the non-treated (or *control*) area (Mississippi).
- The model becomes:

$$math_{it} = \beta_0 + \delta_0 D_{1t} + \beta_1 G_i + \delta_1 G_i D_{1t} + u_{it} \quad t = t_0, t_1, \quad i \in \mathcal{G}, \mathcal{M} \tag{6}$$

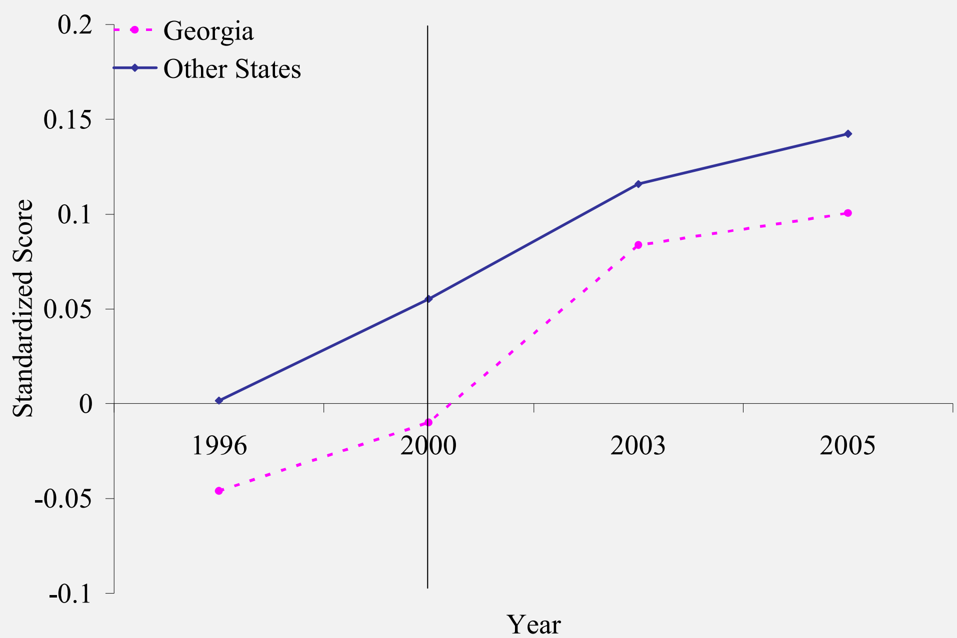
- The OLS estimator for δ_1 becomes:

$$\begin{aligned} \hat{\delta}_1 &= \left(\overline{math}_{G1} - \overline{math}_{M1}\right) - \left(\overline{math}_{G0} - \overline{math}_{M0}\right) & (7) \\ &= \left(\overline{math}_{G1} - \overline{math}_{G0}\right) - \left(\overline{math}_{M1} - \overline{math}_{M0}\right) & (8) \end{aligned}$$

Fitzpatrick (2008)

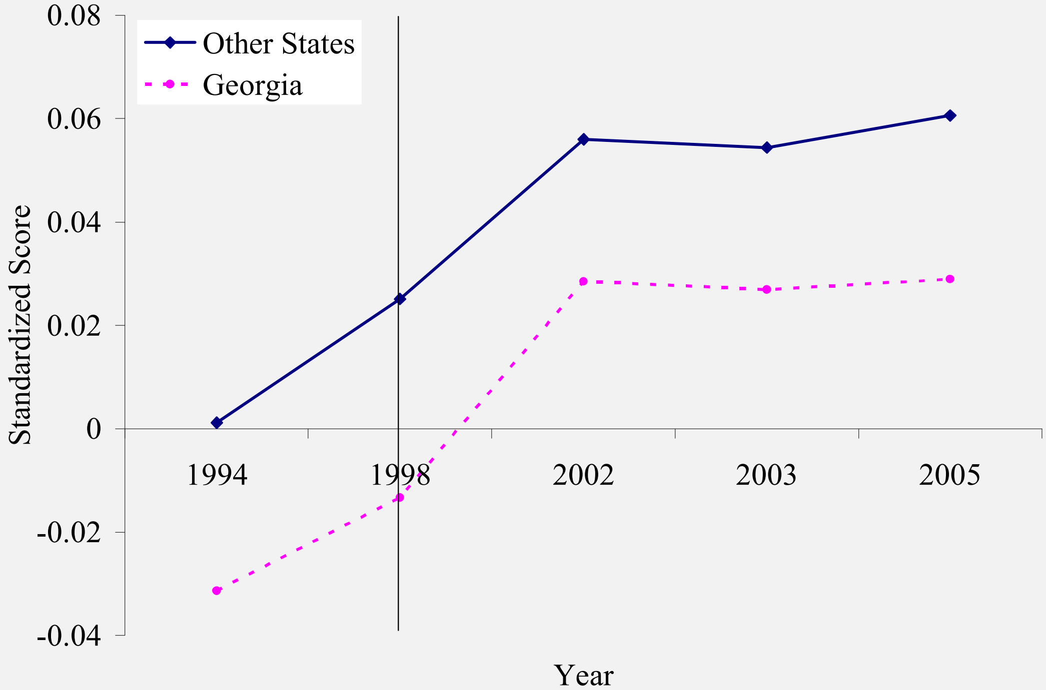
Figure 4. Standardized 4th Grade NAEP Scores, Georgia vs. Rest of the U.S. (Line indicates last pre-program cohort)

Panel A. Mathematics Scores



Fitzpatrick (2008)

Panel B. Reading Scores



Kiel and McCain (1995)

- The treatment here was the construction of an incinerator and the outcome variables of interest was housing price.
- The control group households consisted of those that were not near the incinerator.

	nearinc=1	1981	1978	DID
y81	6,926			18,790
	(8,205)			(4,050)**
nearinc		-30,688	-18,824	-18,824
		(5,828)**	(4,745)**	(4,875)**
y81nrinc				-11,864
				(7,457)
_cons	63,693	101,308	82,517	82,517
	(5,296)**	(3,093)**	(2,654)**	(2,727)**
R^2	0.01	0.17	0.08	0.17
N	96	142	179	321

* $p < 0.05$; ** $p < 0.01$

Two-Period Panel Data Analysis

- We now turn our attention to panel data, where the same units (individuals, firms, counties, etc.) are observed over time.
- We start with the simplest case in which there are only two time periods ($t = 1$ and $t = 2$).
- The generic version of this model would be

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + v_{it} \tag{9}$$

  where $d2_t = 1$ if $t = 2$ (=0 otherwise) and v_{it} is now used to denote the full set of unobserved factors.
- Notice that the model still incorporates shifts in the dependent variable over time (as measured by δ_0).

Pooled OLS

- The simplest approach to estimation in the case of panel data is **Pooled OLS**;

  . . . i.e., stacking the data from the two time periods estimating the model in (9) using OLS.
- The usual assumptions will need to apply in order for consistency and unbiasedness to hold.
- Pooled OLS will not be efficient if their are serial correlations in the residuals.

  . . . the *cluster(id)* option in Stata will make the reported standard error robust to any kind of serial correlation in the errors for a given i and for any form of heteroskedasticity.
- Intertemporal feedback effects can also cause violations of the strict exogeneity assumption.

The Unobserved Heterogeneity

- One of the primary concerns in regression analysis is possibility of omitted variables bias.
- Wooldridge illustrates this problem in the context of a simple regression of 1987 city crime rates on the local unemployment rate:

$$\widehat{crmrta} = \underset{(20.76)}{128.38} \quad \underset{(3.42)}{-4.16}\, unem$$

- This counter-intuitive result is likely due to omitted factors, such as
  - law enforcement expenditures;
  - cultural forces;
  - age distribution;
  - local economic conditions, etc.
- A panel structure will allow us to control for some of these unobserved factors, specifically those that are constant over time.

The Unobserved Heterogeneity (cont'd)

- To see this, we decompose the error term into two components,

$$v_{it} = a_i + u_{it} \tag{10}$$

  where a_i captures all *time constant* unobserved factors, u_{it} denotes unobserved (*idiosyncratic*) factors that change over time.
- The term a_i is referred to in the literature by a number of names, including
  - Unobserved heterogeneity;
  - Individual fixed effects, firm fixed effects, etc.
  - Unobserved effect.

The Unobserved Heterogeneity (cont'd)

- Panel data allows us to eliminate the impact of the unobserved heterogeneity through first differencing.
- Specifically, writing out the model in (9) for the two time periods, we get

$$y_{i2} = \beta_0 + \delta_0 \quad + \beta_1 x_{i21} + \cdots + \beta_k x_{i2k} + a_i + u_{i2} \tag{11}$$

$$y_{i1} = \beta_0 \quad + \beta_1 x_{i11} + \cdots + \beta_k x_{i1k} + a_i + u_{i1} \tag{12}$$

- Subtracting (12) from (11) yields

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \cdots + \beta_k \Delta x_{ik} + \Delta u_i \tag{13}$$

where $\Delta y_i = y_{i2} - y_{i1}$, $\Delta x_{ij} = x_{i2j} - x_{i1j}$ $j = 1, \ldots, k$, and $\Delta u_i = u_{i2} - u_{i1}$.
- Note that a_i no longer appears in the model.

The Tradeoffs

- The advantage of the so-called **first differencing estimator** is that, *if* the model in (13) satisfies the usual assumptions, OLS will be
  - *BLUE* (Theorem 3.4) or consistent (Theorem 5.1);
  - purged of the potential omitted variables biased induced by the unobserved heterogeneity.
- There are, however, costs:
  1. For OLS to be consistent, we need $E(\Delta u_i | \Delta \boldsymbol{x}_i) = 0$, which effectively requires $E(u_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}) = 0, t = 1, 2$
     ...which can be violated when there is inter-temporal feedback;
  2. We can no longer estimate parameters associated with observed variables that are constant over time (e.g., regional dummies, race,) or increase at a constant rate over time (e.g., age);
  3. Even if a variable varies over time (e.g., education), the variation may be limited, making for imprecise parameter estimates (see Example 13.5 in Wooldridge);
  4. Measurement error can be a more pronounced problem.

Example: Crime Rate and Unemployment

Crime Rate Panel

	1987	1982/87	First Diff
Dependent Variable	$crmrte_{it}$	$crmrte_{it}$	$\Delta crmrte_i$
$unem_{it}$	-4.16	0.43	
	(3.42)	(1.19)	
$d87_{it}$		7.94	
		(7.98)	
$\Delta unem_i$			2.22
			(0.88)*
_cons	128.38	93.42	15.40
	(20.76)**	(12.74)**	(4.70)**
R^2	0.03	0.01	0.13
N	46	92	46

* $p < 0.05$; ** $p < 0.01$

Policy Analysis with Two-Period Panel Data

- In the earlier part of this chapter, we used the *Diff-in-Diff* estimator to control for *cohort* fixed effects when we had repeated cross-sections.
- A similar approach can be used when we have panel data ...only now we can control for individual level heterogeneity.
- The model takes the form:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it} \tag{14}$$

- The term a_i plays a role similar to our treatment group indicator in the *Diff-in-Diff* estimator.
- Note: In the *Diff-in-Diff* case, the program was in effect only for the treatment group and then only during the second period.
- Here, all we need is that $prog_{it}$ varies over time for some or all of the observations. (See Example 13.7 in Wooldridge using drunk driving laws).

Differencing with More than Two Time Periods

- With more than two time periods in a panel data set, first differencing can be applied as well.
- Suppose we have T time periods. Then our model takes the form

$$_{it} = \beta_0 + \delta_2 d2_t + \cdot + \delta_T dT_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it} \quad (15)$$

  where $ds_t = 1$ if $t = s$ (=0 otherwise) for $s = 2, \ldots, T$.
- For each period $t = 2, \ldots, T$ we can first differenced adjacent periods to yield the model

$$\Delta y_{it} = \delta_2 \Delta d2_t + \cdot + \delta_T \Delta dT_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it} \quad (16)$$

  where now $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta x_{itj} = x_{itj} - x_{i,t-1,j}$ $j = 1, \ldots, k$, and $\Delta u_{it} = u_{it} - u_{i,t-1}$.
- The model in (16) has $n(T-1)$ observations.

Estimation of the FD Model

- Employing the usual assumptions, (16) can be estimated using **Pooled OLS**.

  . . . This involves stacking the data from the different time periods and running OLS.
- The differenced errors, however, are potentially serially correlated, though corrections exist (See Wooldridge, 2010, Chapter 10).
- The first difference estimator with $T > 2$ will, of course, suffer from the same potential problems noted for the $T = 2$ case.