41903: Basic Panel Data Methods

Notes 3
Differences-in-Differences

Simple Differences-in-Differences (DD):
- observe the outcome of individuals in two groups, Control (C) and Treatment (T)
- observe outcome at two different time periods, Pre (0) and Post (1)
- treatment group receives treatment between times Pre and Post (i.e. pre-treatment outcome is measured at time 0 and post-treatment outcome is measured at time 1)
- **Goal:** Estimate causal effect of treatment

E.g.
1. **Health intervention**
   - observe health outcomes in two villages at the beginning of 2010
   - a sanitary water supply is provided in the first village during 2010
   - health outcomes are observed in the two villages at the beginning of 2011

2. **Public policy**
   - observe employment in two states at the beginning of 2010
   - the minimum wage is increased in the first state during the spring of 2010
   - observe employment outcomes in the same two states in 2011
Estimator of causal effect:

\[ \hat{\beta}_{DD} = (\bar{y}_{T1} - \bar{y}_{T0}) - (\bar{y}_{C1} - \bar{y}_{C0}) \]

- \( \bar{y}_{g,t} \) is the sample average of the outcome for group \( g \) at time \( t \)

Expected value of DD estimator:

\[
E[\hat{\beta}_{DD}] = (E[\bar{y}_{T1}] - E[\bar{y}_{T0}]) - (E[\bar{y}_{C1}] - E[\bar{y}_{C0}])
= (\beta_{TE} + \delta_{T1} + \alpha_{T} - \alpha_{T}) - (\delta_{C1} + \alpha_{C} - \alpha_{C})
= \beta_{TE} + (\delta_{T1} - \delta_{C1})
\]

- \( \beta_{TE} \) = treatment effect
- \( \alpha_{T}, \alpha_{C} \) = baseline averages in groups \( T \) and \( C \)
- \( \delta_{C1} \) = change in average outcomes in group \( C \) between times 0 and 1
- \( \delta_{T1} \) = change in average outcomes in group \( T \) in the absence of treatment
Common Trends Assumption

Have

\[ E[\hat{\beta}_{DD}] = \beta_{TE} + (\delta_{T1} - \delta_{C1}) \]

**Common trends assumption:** \( \delta_{C1} = \delta_{T1} \)

- \( \Rightarrow E[\hat{\beta}_{DD}] = \beta_{TE} \)
- \( \hat{\beta}_{DD} \) is unbiased for the treatment effect under assumption that the two groups would have trended the same in the absence of treatment
- allows for quite general initial difference between the groups (that can be correlated to treatment receipt)
Diffs-in-Diffs as Linear Regression

Regression model:

\[ y_{it} = \alpha + \alpha_T T_{it} + \delta D_{it} + \beta T_{it}D_{it} + \varepsilon_{it} \]

- \( T_{it} = 1 \) if \( i \in T \) (treatment group dummy)
- \( D_{it} = 1 \) if \( t = 1 \) (post-treatment dummy)

Model implies

1. \( E[y_{it}|i \in C, t = 0] = \alpha \)
2. \( E[y_{it}|i \in C, t = 1] = \alpha + \delta \)
3. \( E[y_{it}|i \in T, t = 0] = \alpha + \alpha_T \)
4. \( E[y_{it}|i \in T, t = 1] = \alpha + \alpha_T + \delta + \beta \)

\[ 1.-4. \Rightarrow \beta = (E[y_{it}|i \in T, t = 1] - E[y_{it}|i \in T, t = 0]) - (E[y_{it}|i \in C, t = 1] - E[y_{it}|i \in C, t = 0]) = \beta_{TE} \]

- Also immediate that \( \hat{\beta}_{OLS} = \hat{\beta}_{DD} \) in this model.
  - Facilitates inference
Richardson and Troost (2009) - examine effect of monetary policy on bank failure during Great Depression

Story:

- Caldwell ran largest banking chain in south in 1920’s. Failed with stock market crash in October 1929. “Caldwell Collapse” precipitated bank runs in South in December 1930
- Border between 6th (Atlanta Fed) and 8th (St. Louis Fed) Federal Reserve Districts cuts through the middle of Mississippi
- Atlanta Fed favored lending to troubled banks - Fed increased lending by 40% within four weeks of Caldwell crisis
- St. Louis Fed held that lending should be restricted during a recession - Fed decreased lending by 10% within four weeks of Caldwell crisis
- Richardson and Troost argue 8th district is control (policy effectively to do nothing); argue 6th district is treatment (policy to increase lending)
Monetary Policy DD: Some Numbers

- 121 banks in operation in 1931 in MS district 6 (after Caldwell Crisis)
- 135 banks in operation in 1930 in MS district 6 (before Caldwell Crisis)
- 132 banks in operation in 1931 in MS district 8 (after Caldwell Crisis)
- 165 banks in operation in 1930 in MS district 8 (before Caldwell Crisis)

Naive estimate of treatment effect:

\[ y_{treat} - y_{control} = \text{banks in business}_{Post,6} - \text{banks in business}_{Post,8} = 121 - 132 = -11 \]

Ignores baseline differences

Diffs-in-Diffs estimate:

\[ (y_{treat, post} - y_{treat, pre}) - (y_{control, post} - y_{control, pre}) = (121 - 135) - (132 - 165) = 19 \]

Suggests 19 banks saved by looser monetary policy.
Counterfactual obtained assuming trend would have been the same in the treatment and control group in the absence of treatment.
Monetary Policy DD: DD Regression

\[ \text{banks in business}_{it} = \alpha + \alpha_T \text{Region 6}_{it} + \delta \text{After 1930}_{it} + \beta \text{Region 6}_{it} \text{After 1930}_{it} + \varepsilon_{it} \]

Results using 1930 and 1931 (estimated s.e. in parentheses):

- \( \hat{\alpha} = 165 \) (. )
- \( \hat{\alpha}_T = -30 \) (. )
- \( \hat{\delta} = -33 \) (. )
- \( \hat{\beta} = 19 \) (. )

Why are the s.e. all missing?
Nice thing about DD regression - immediate how to generalize to more time periods/groups

\[ banks \text{ in business}_{it} = \alpha + \alpha_T \text{Region 6}_{it} + \delta \text{After 1930}_{it} + \beta \text{Region 6}_{it} \text{After 1930}_{it} + \epsilon_{it} \]

Results using all data 1929 and 1934 (estimated s.e. in parentheses):

- $\hat{\alpha} = 167 (6.19)$
- $\hat{\alpha}_T = -29 (8.75)$
- $\hat{\delta} = -49 (7.58)$
- $\hat{\beta} = 20.5 (10.72)$

IID, homoscedastic standard errors
Monetary Policy DD: DD Regression Picture

The graph shows the number of banks in business for different regions over the years 1929 to 1934. The number of banks in Region 8 and Region 6 decreased steadily over the years. The graph also includes a counterfactual line for Region 6, which remains relatively flat compared to the other regions.
Augmenting the DD Model

Easy to extend DD model in variety of ways

1) Common trends after controlling (linearly) for exogenous variables $x_{it}$:

$$y_{it} = \alpha + \alpha_T T_{it} + \delta D_{it} + \beta T_{it} D_{it} + x_{it}' \gamma + \varepsilon_{it}$$

2) Multiple treatment and control groups ($M$ total), effect of treatment homogeneous across groups, common trends after controlling linearly for exogenous variables $x_{it}$:

$$y_{it} = \sum_{j=1}^{M} \alpha_j d_{i \in j} + \delta D_{it} + \beta T_{it} D_{it} + x_{it}' \gamma + \varepsilon_{it}$$

where $d_{i \in j}$ is a dummy which is one if observation $i$ belongs to group $j$. 

Basic Panel Data
Augmenting the DD Model

3) Multiple time periods \((0, 1, ..., T)\), treatment may occur at different times for different groups, more than two cross-sectional units \((M\) in total), effect of treatment homogeneous across groups, common trends after controlling linearly for exogenous variables \(x_{it}\):

\[
y_{it} = \sum_{j=1}^{M} \alpha_j d_{i \in j} + \sum_{s=1}^{T} \delta_s D_{is} + \beta \tilde{T}_{it} + x_{it}' \gamma + \varepsilon_{it}
\]

where \(D_{is}\) is a dummy equal to one if \(t = s\) and \(\tilde{T}_{it}\) is a dummy which equals one whenever \(\{i, t\}\) corresponds to an individual who was treated at time \(t\).

Same as a “fixed effects” panel data model where \(\alpha_j\) denotes time-invariant heterogeneity for “individual \(j\)” (individual effects) and \(\delta_s\) are cross-sectionally invariant heterogeneity (time effects)

Note that common trends assumption is parameterization dependent: e.g. common trends in levels implies no common trends in logs

- Athey and Imbens (2006) propose a nonparametric extension that is invariant to scaling of the outcome
Example: Health Insurance Reform

Kaestner and Simon (ILLR 2002)

Individual level data from CPS March Supplement

Data:

- 36999 individuals in 51 states (including DC) and 10 years
- Outcome: \( \log(\text{wage}) \)
- Variables of Interest (Treatments):
  - \( \text{fullr} \) - “full” HI reform indicator
  - \( \text{partialr} \) - “partial” HI reform indicator
  - \( \text{highcost} \) - number of mandates to cover high cost procedures
  - \( \text{women} \) - number of mandates for covering largely female specific procedures
  - \( \text{other} \) - number of mandates covering other procedures
- Controls:
  - \( \text{age} \): dummies for age categories \((a2 - a8)\)
  - \( \text{gender} \): dummy for male
  - \( \text{education} \): dummies for high school, some college, and college
  - \( \text{maritalstatus} \): dummies for married and divorced
  - \( \text{race} \): dummies for African American and other race (white excluded category)
  - \( \text{children} \): number of children < 6 and number of children between 6 and 18
  - state effects, year effects, state specific trends
HI Reform Regression

Estimated model:

$$\log(wage)_{it} = \alpha_s + \kappa_t + \delta_s t + treatment_{st}^\prime \gamma + x_{it}^\prime \beta + \varepsilon_{it}$$

where $treatment_{st}$ is a $5 \times 1$ vector of the state level treatment variables applying to individual $i$ at time $t$.

- Note the presence of state effects, time effects, and state-specific linear trends
- Multiple groups, time periods, and treatment variables that vary across both dimensions allow us to relax common trend assumption
- Can only learn treatment effect from variation in data that is orthogonal to initial (or average) state-level features, an arbitrarily flexible national trend, and state-specific linear trends
- I.e. common trends relaxed to common trends not explained by initial conditions, general national trend, state-specific linear trends

Should definitely use clustered standard errors or another similar method.
HI Reform Regression: Results

Estimation Results:

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}_{DD}$</th>
<th>IID</th>
<th>Cluster(State)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Reform</td>
<td>-0.0332</td>
<td>0.0150</td>
<td>0.0212</td>
</tr>
<tr>
<td>Partial Reform</td>
<td>0.0036</td>
<td>0.0143</td>
<td>0.0168</td>
</tr>
<tr>
<td># High Cost</td>
<td>-0.0021</td>
<td>0.0087</td>
<td>0.0072</td>
</tr>
<tr>
<td># Women</td>
<td>0.0065</td>
<td>0.0077</td>
<td>0.0075</td>
</tr>
<tr>
<td># Other</td>
<td>0.0022</td>
<td>0.0058</td>
<td>0.0073</td>
</tr>
</tbody>
</table>

Note: Bonferroni critical value (5%) $\approx 2.68$
Linear Unobserved Effects Model

Most commonly specified panel model in economics:

\[ y_{it} = x_{it}' \beta + \alpha_i + \delta_t + \epsilon_{it} \]

- \( i = 1, \ldots, N, \ t = 1, \ldots, T \)
  - ignoring unbalanced panels
  - no substantive implications (though annoying notationally) as long as observations missing at random
- \( \beta \) are the parameters of interest
- \( \alpha_i \) and \( \delta_t \) are unobserved time invariant (individual) and cross-sectionally invariant (time) effects
  - nuisance parameters

Matrix notation:

\[ Y = X\beta + D\alpha + R\delta + \epsilon \]

- \( Y \) is \( NT \times 1 \); \( X \) is \( NT \times k \); \( \epsilon \) is an \( NT \times 1 \) vector
- \( D \) is an \( NT \times N \) matrix of dummy variables for each individual
- \( \alpha \) is the \( N \times 1 \) vector \( (\alpha_1, \alpha_2, \ldots, \alpha_N)' \)
- \( R \) is an \( NT \times T \) matrix of dummy variables for each time period
- \( \delta \) is the \( T \times 1 \) vector \( (\delta_1, \delta_2, \ldots, \delta_T)' \)
Some Technicalities

Traditional panel data analysis

- $N \gg T$ with $N \to \infty$ and $T$ fixed
- Take $\alpha$ and $\delta$ as unrestricted nuisance parameters ("fixed effects") that need to be estimated or eliminated
- $\delta_t$ are included in $\beta$ - $X$ is redefined as $[X, R]$ and $\beta$ is redefined as $(\beta', \delta')'$

Focus on model

$$Y = X\beta + D\alpha + \varepsilon$$

where time effects are implicitly allowed for in the $X$ matrix.

Normalization:

- $\alpha$, $\delta$, and an intercept cannot be jointly identified - need a normalization
- Use intercept $= 0$ and one element of $\delta = 0$ normalization in following - innocuous relative to other potential normalizations
- $T \gg N$ roughly symmetric switching $\alpha$ and $\delta$
- $T \sim N$, should really explicitly consider both $\alpha$ and $\delta$
- “Random effects” approaches place restrictions on unobserved heterogeneity
  - May be very useful in nonlinear settings or settings with more complicated unobserved heterogeneity (e.g. interactive effects)
  - Not particularly popular in empirical economics
Estimation

Estimate $\beta$ and $\alpha$ using OLS:

$$(\hat{\beta}', \hat{\alpha}')' = (W'W)^{-1}(W'Y)$$

- $W = [X \ D]$; assume $W$ full rank
- Sometimes referred to as LSDV regression, but usually just fixed effects

Usually don’t care about $\alpha$, focus on $\beta$:

$$\hat{\beta} = (X' M_D X)^{-1}(X' M_D Y), \quad M_D = \mathbf{I}_{NT} - D(D'D)^{-1} D'$$

$M_DA$ for conformable vector $A$ has simple form:

$$[M_D A]_{i,t} = a_{it} - \frac{1}{T} \sum_{s=1}^{T} a_{is} = a_{it} - \bar{a}_i$$

- LSDV/fixed effects estimator equivalent to regression of $y_{it} - \bar{y}_i$ on $x_{it} - \bar{x}_i$
- One more name: Within group estimator - $\beta$ estimated using only information about how data deviate from within group means
  - within group transformation explicitly accounting for state and time effects: $a_{i,t} - \bar{a}_i - \bar{a}_t + \bar{a}$ (needs balanced panel)
We get wildly different estimates of the slope depending on whether we pool or use the within group variation!!!
Conventional Asymptotic Analysis

FE estimator:

\[ \hat{\beta}_{FE} = \beta + \left( \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \left( \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i) \right) \]

Clear get good properties under

a. \( \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \) satisfies an LLN and is full rank

b. \( \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i) \overset{p}{\rightarrow} 0 \)
   ▶ Note that the usual orthogonality condition \( E[\varepsilon_{it}x_{it}] = 0 \) is not sufficient to guarantee that c is satisfied.
   ▶ A common sufficient condition is “mean independence”: \( E[\varepsilon_{it}|\alpha_i, x_{i1}, ..., x_{iT}] = 0 \)
   ▶ A weaker sufficient condition is \( E[\varepsilon_{it}x_{is}] = 0 \) for all \( s = 1, ..., T, t = 1, ..., T \)
   ▶ Need some form of strict exogeneity

c. \( \frac{1}{\sqrt{NT}} \sum_{i,t} (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i) \) satisfies a CLT
   ▶ If \( \{x_{it}, \varepsilon_{it}\} \) independent across \( i \), can use CLT for independent data \( A_i \) since
   \[ \frac{1}{\sqrt{NT}} \sum_{i,t} (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i) = \frac{1}{\sqrt{N}} \sum_i A_i \] where \( A_i = \frac{1}{\sqrt{T}} (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i) \) are independent.
   ▶ If the data are cross-sectionally correlated, need an appropriate spatial CLT (e.g. Jenish and Prucha (2007))
Assume $\varepsilon_{it}$ iid, independent of $x_i$, $E[\varepsilon_{it}^2] = \sigma^2$.

FE is Gauss-Markov estimator; FE is BLUE.

Easy to show that $V = Q^{-1} \Omega Q^{-1} = \sigma^2 Q^{-1}$ for $Q = E[(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)']$

Use $\hat{V} = \hat{\sigma}^2 \hat{Q}^{-1}$

- $\hat{Q} = \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$
- $\hat{\sigma}^2 = \frac{1}{NT-N-k} \sum_{i,t} e_{it}^2$ where $e_{it} = y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)' \hat{\beta}_{FE}$ and $k = \text{dim}(x_{it})$
Assume observations independent across $i$.

Use $\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}$ where $\hat{Q} = \frac{1}{NT} \sum_{i,t}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$ as above.

In this case,

$$\Omega = \lim_{N \to \infty} \text{Var}\left(\frac{1}{\sqrt{NT}} \sum_{i,t}(x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i)\right)$$

$$= \lim_{N \to \infty} \frac{1}{N} \sum_{i} \text{Var}\left(\frac{1}{\sqrt{T}} \sum_{t}(x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i)\right)$$

Most widely used estimator of $\Omega$ is “Clustered Covariance Estimator”:

$$\hat{\Omega} = \frac{1}{N} \sum_{i} \frac{1}{T}(\tilde{X}_i' E_i)(\tilde{X}_i' E_i)'$$

where $\tilde{X}_i$ is the $T \times k$ matrix obtained by stacking $(x_{it} - \bar{x}_i)'$ for each individual and $E_i = (e_{i1}, e_{i2}, ..., e_{iT})'$ for $e_{it}$ defined above.

Allows for arbitrary within group heteroskedasticity and serial correlation.
(First) Differencing

Can take differences to remove time invariant heterogeneity in the linear model

Model in first-differences:

\[ y_{it} - y_{it-1} = (x_{it} - x_{it-1})' \beta + \epsilon_{it} - \epsilon_{it-1} \]

First-difference estimator:

\[
\hat{\beta}_{FD} = \left( \frac{1}{N(T-1)} \sum_{i} \sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)^{-1} \left( \frac{1}{N(T-1)} \sum_{i} \sum_{t=2}^{T} \Delta x_{it} \Delta y_{it} \right)
\]

- With \( T = 2 \), FD equivalent to FE (differ otherwise)
- Can take other differences
- If \( \epsilon_{it} = \epsilon_{it-1} + \nu_{it} \) with \( \nu_{it} \) iid (i.e. is a random walk), FD is BLUE
- Differencing will only remove time invariant FE in linear model
  - Transformations to remove FE available only in a small number of models (e.g. linear, logit, poisson)
FD Asymptotic Properties

Analysis pretty similar to FE and other LS estimators

- Estimator is consistent and asymptotically normal under the usual types of conditions
- s.e. estimation pretty much the same as for FE

Key difference is in orthogonality conditions

- “mean independence” is sufficient: $E[\varepsilon_{it} | \alpha_i, x_{it1}, ..., x_{iT}] = 0$ (just as in FE)
- A weaker sufficient condition is $E[\varepsilon_{it} x_{it}] = E[\varepsilon_{it} x_{it-1}] = E[\varepsilon_{it-1} x_{it}] = 0$ for all $t = 1, ..., T$.
  - less stringent than “weak” orthogonality condition from FE
- Not considering large $T$ case explicitly, but note that $E[x_{it} \varepsilon_{it}] = 0$ would be sufficient for FE with large $T$ while still need $E[\varepsilon_{it} x_{it}] = E[\varepsilon_{it} x_{it-1}] = E[\varepsilon_{it-1} x_{it}] = 0$ for FD.
District level data similar to school level data in Papke (2005).

Data:

- Annual district level data on Michigan school districts from 1992-1998
- Outcome: \(\text{math4}\) Fraction of 4th grade students receiving a passing score on a standardized math test
- Variable of Interest: \(\log(rexpp)\) Real expenditures (1997) dollars per pupil in the district
- Controls: \(\log(enrol)\) district level enrollment, \(lunch\) fraction of students in the district eligible for free lunch program
- 550 districts over 7 years (grouped in 57 “Intermediate School Districts”)

Goal: Estimate causal effect of school expenditures on student outcomes (measured by math scores)
Test scores: Results

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}_{POLS}$</th>
<th>IID</th>
<th>Clus</th>
<th>$\hat{\beta}_{FE}$</th>
<th>IID</th>
<th>Clus</th>
<th>$\hat{\beta}_{FD}$</th>
<th>IID</th>
<th>Clus</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.Expend</td>
<td>15.01</td>
<td>2.57</td>
<td>2.81</td>
<td>-0.411</td>
<td>2.243</td>
<td>3.065</td>
<td>-1.41</td>
<td>2.77</td>
<td>5.34</td>
</tr>
<tr>
<td>I.Expend$_{t-1}$</td>
<td>15.65</td>
<td>2.42</td>
<td>2.75</td>
<td>7.00</td>
<td>2.16</td>
<td>4.21</td>
<td>11.04</td>
<td>2.54</td>
<td>5.26</td>
</tr>
<tr>
<td>I.Enroll</td>
<td>-0.221</td>
<td>0.233</td>
<td>0.617</td>
<td>0.245</td>
<td>1.004</td>
<td>0.914</td>
<td>2.14</td>
<td>1.07</td>
<td>1.72</td>
</tr>
<tr>
<td>Lunch</td>
<td>-0.389</td>
<td>0.015</td>
<td>0.047</td>
<td>0.062</td>
<td>0.047</td>
<td>0.127</td>
<td>0.073</td>
<td>0.056</td>
<td>0.149</td>
</tr>
</tbody>
</table>

- Clustering at intermediate school level
- $\hat{\beta}_{POLS}$ is pooled OLS estimator
- $\hat{\beta}_{FE}$ is fixed effects estimator
- $\hat{\beta}_{FD}$ is first difference estimator

Basic Panel Data
Hausman Test Results

FD and FE look reasonably close.

Hausman test comparing coefficients of interest (coefficients on log(expenditure) and L.log(expenditure)).
  - $H = 4.096$ (distributed as $\chi^2(2)$ under the null that both are estimating the same thing)
  - p-value: .129

Hausman test comparing all 4 district level coefficients:
  - $H = 4.619$ (distributed as $\chi^2(4)$ under the null that both are estimating the same thing)
  - p-value: .329
Robustness check: Putting in a lead

What happens if we put in a lead of log(Expenditure) (i.e. log(\(r_{e x p p_{i t+1}}\)))?

Fixed Effects Estimates including Lead:

<table>
<thead>
<tr>
<th></th>
<th>(\hat{\beta}_{FE})</th>
<th>IID</th>
<th>Clus</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Expenditures)</td>
<td>-2.19</td>
<td>2.18</td>
<td>3.82</td>
</tr>
<tr>
<td>L.log(Expenditures)</td>
<td>6.21</td>
<td>2.10</td>
<td>4.93</td>
</tr>
<tr>
<td>F.log(Expenditures)</td>
<td>-1.28</td>
<td>2.26</td>
<td>4.39</td>
</tr>
</tbody>
</table>

First-Difference Estimates including Lead:

<table>
<thead>
<tr>
<th></th>
<th>(\hat{\beta}_{FD})</th>
<th>IID</th>
<th>Clus</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Expenditures)</td>
<td>-3.35</td>
<td>2.68</td>
<td>4.88</td>
</tr>
<tr>
<td>L.log(Expenditures)</td>
<td>10.18</td>
<td>2.43</td>
<td>5.23</td>
</tr>
<tr>
<td>F.log(Expenditures)</td>
<td>0.66</td>
<td>2.59</td>
<td>4.83</td>
</tr>
</tbody>
</table>
A **predetermined regressor** satisfies $E[x_{it-j} \varepsilon_{it}] = 0$ for all $j \geq 0$ but may have $E[x_{it+j} \varepsilon_{it}] \neq 0$ for $j > 0$.

Most common example is a lagged dependent variable - $y_{it-1}$

Example: Dynamic Panel Model

$$y_{it} = \rho y_{it-1} + \alpha_i + \varepsilon_{it}$$

- $\varepsilon_{it}$ iid
- Assume $E[y_{it-j} \varepsilon_{it}] = 0$ for all $j \geq 0$
- Clear that $E[y_{it-j} \varepsilon_{it-j}] \neq 0$ for any $j$
- Implies that both FE and FD orthogonality conditions are violated
Generically, fixed effects estimators in nonlinear models and models without strict exogeneity are strongly biased unless $T$ is very large

- Fixed effects are estimated from $T$ observations
  - Will have small sample bias that shrinks like $1/T$ in nonlinear models or with predetermined regressors

- Common parameters are estimated from $NT$ observations
  - Have variance that shrinks like $1/NT$
  - Inherit $1/T$ bias from fixed effects - Variance is decreasing much faster than bias

- This feature of fixed effects estimators is termed the incidental parameters problem
  - Noted in Neyman and Scott (1949) in context of Gaussian ML model focused on estimating common common variance
FE Dynamic Panel Model (1)

\[ y_{it} = \rho y_{it-1} + \alpha_i + \varepsilon_{it} \]

Concentrate out the \( \alpha_i \):

\[ \hat{\alpha}_i(\rho) = \arg \min_a \sum_{t=1}^{T} (y_{it} - \rho y_{it-1} - a)^2 \]

\[ = \frac{1}{T} \sum_{t=1}^{T} (y_{it} - \rho y_{it-1}) \]

\[ = \bar{y}_i - \rho \bar{y}_{i,-1} \]

- Have \( T + 1 \geq 3 \) observations for \( 0 \leq t \leq T \) (by convention)
- \( \bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}, \quad \bar{y}_{i,-1} = \frac{1}{T} \sum_{t=0}^{T-1} y_{it} = \frac{1}{T} \sum_{t=1}^{T} y_{it-1} \)

Problem we need to solve for \( \rho \):

\[ \hat{\rho} = \arg \min_r \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} ((y_{it} - \bar{y}_i) - r(y_{it-1} - \bar{y}_{i,-1}))^2 \]
FE Dynamic Panel Model (2)

FOC for $\rho$:

$$S(r) = \frac{-2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} ((y_{it} - \bar{y}_i) - r(y_{it-1} - \bar{y}_{i,-1}))(y_{it-1} - \bar{y}_{i,-1})$$

Important thing to see

$$E[S(r)|r = \rho] = \frac{-2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E[(\varepsilon_{it} - \bar{\varepsilon}_{it})(y_{it-1} - \bar{y}_{i,-1})]$$

$$= \frac{-2}{T} \sum_{t=1}^{T} [0 - E[\bar{\varepsilon}_i y_{it-1}] - E[\varepsilon_{it} \bar{y}_{i,-1}] + E[\varepsilon_{it} \bar{y}_{i,-1}]]$$

$$\neq 0$$

because $y_{it-1}$ is not strictly exogenous
Exact expressions available (assuming $|\rho| < 1$ and $\varepsilon_{it}$ iid) by noting

$$y_{it} = \frac{\alpha_i}{1 - \rho} + \sum_{j=0}^{\infty} \rho^j \varepsilon_{it-j}$$

and doing some tedious algebra.

$$E[S(r)|r = \rho] = -2 \frac{\sigma_{\varepsilon}^2}{T(1 - \rho)} \sum_{t=1}^{T} \left( 1 - \rho^{t-1} - \rho^{T-t} + \frac{1}{T} \frac{1 - \rho^T}{1 - \rho} \right)$$

Expected score for parameter of interest not equal to 0 in population - Estimator sets sample average of score equal to 0.

Not solving the correct moment condition in sample $\Rightarrow$ inconsistent estimator
FE Dynamic Panel Model Estimator

\[ \hat{\rho} = \rho + \frac{1}{N_T} \sum_{i=1}^{N} \sum_{t=1}^{T} (\varepsilon_{it} - \bar{\varepsilon}_i)(y_{it-1} - \bar{y}_{i,-1}) \]

where:

\[ \frac{1}{N_T} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it-1} - \bar{y}_{i,-1})^2 \]

Take limits as \( N \to \infty \)

\[ \hat{\rho} - \rho \xrightarrow{p} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[(\varepsilon_{it} - \bar{\varepsilon}_i)(y_{it-1} - \bar{y}_{i,-1})] \]

\[ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[(y_{it-1} - \bar{y}_{i,-1})^2] \]

\[ = -\frac{(1 + \rho)}{T - 1} \left[ 1 \right. \right.

\[ - \frac{1}{T} \left( 1 - \frac{(1 - \rho^T)}{1 - \rho} \right) \times \]

\[ \left. \left. \left[ 1 - \frac{2\rho}{(1 - \rho)(T - 1)} \left( 1 - \frac{1}{T} \left( 1 - \rho^T \right) \right) \right]^{-1} \right. \]

\[ = \frac{b_{\frac{T}{T}}}{T} \]

Estimator is inconsistent as \( N \to \infty \).
Note

\[ b_T^* = -(1 + \rho) \frac{T}{T - 1} + O \left( \frac{1}{T} \right) = -(1 + \rho) + O \left( \frac{1}{T} \right) = b + O \left( \frac{1}{T} \right) \]

Consistency (can be made formal):

- asymptotic bias \[ \frac{b_T^*}{T} = \frac{b}{T} + O \left( \frac{1}{T^2} \right) \rightarrow 0 \text{ as } T \rightarrow \infty \]
- \[ \hat{\rho} \xrightarrow{p} \rho \text{ as } (N, T) \rightarrow \infty \]

Asymptotic Normality (can be made formal):

\[ \sqrt{NT} (\hat{\rho} - \rho) = N(0, V) + b \sqrt{\frac{N}{T}} + O \left( \frac{\sqrt{N}}{T^{3/2}} \right) + o_p(1) \]

- Diverges if \( \frac{N}{T} \rightarrow \infty \)
- Converges to \( N(\tau b, V) \) if \( \frac{N}{T} \rightarrow \tau, \ 0 < \tau < \infty \)
- Converges to \( N(0, V) \) if \( \frac{N}{T} \rightarrow 0 \)
With $T \to \infty$, can remove incidental parameter bias by bias correction

In dynamic panel model, estimate bias $\hat{b} = -(1 - \hat{\rho})$

Bias corrected estimator

$$\hat{\rho} = \hat{\rho} - \frac{\hat{b}}{T}$$

Asymptotic Normality (can be made formal):

$$\sqrt{NT}(\hat{\rho} - \rho) = N(0, V) + (b - \hat{b})\sqrt{\frac{N}{T}} + O\left(\frac{\sqrt{N}}{T^{3/2}}\right) + o_p(1) = N(0, V) + \sqrt{T}(b - \hat{b})\sqrt{\frac{N}{T^3}} + O\left(\sqrt{\frac{N}{T^3}}\right) + o_p(1) = N(0, V) + O_p\left(\sqrt{\frac{N}{T^3}}\right)$$

- Converges to $N(0, V)$ if $\frac{N}{T^3} \to 0$

All of this is formalized in Hahn and Kuersteiner (2003)
Anderson and Hsiao (1981, 1982); Arellano and Bond (1991) look at model in first-differences:

\[ y_{it} - y_{it-1} = \rho(y_{it-1} - y_{it-2}) + (\varepsilon_{it} - \varepsilon_{it-1}) \]

- FE has been eliminated
- RHS variable \( \Delta y_{it-1} = y_{it-1} - y_{it-2} \) correlated to error term \( \Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1} \) because \( y_{it-1} \) correlated to \( \varepsilon_{it-1} \)
- Under assumption that \( \mathbb{E}[\varepsilon_{it} | \mathcal{I}_{t-1}] = 0 \), any \( t - 2 \) lagged variables are potential instruments
- Lots of variants depending on how lagged information is used
- Have weak identification if data are highly persistent
- Have to be comfortable with using lags as instruments
Look at dynamic model of airfare

- 1149 airline routes, 3 years of data
- outcome - $\log(\text{airfare})$
- RHS variables
  - first lag of $\log(\text{airfare})$
  - $\text{concen}$ - fraction of market share accounted for by biggest provider (taken as exogenous)
  - time dummies

Model:

$$\log(\text{airfare}_{it}) = \rho \log(\text{airfare}_{it-1}) + \beta \text{concen}_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$
### Results

<table>
<thead>
<tr>
<th></th>
<th>POLS</th>
<th>FE</th>
<th>FE-BC</th>
<th>FD-IV</th>
<th>FD-ABIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>L.log(airfare)</td>
<td>.9298</td>
<td>.0774</td>
<td>.5387</td>
<td>.3744</td>
<td>.3326</td>
</tr>
<tr>
<td></td>
<td>.0047</td>
<td>.0319</td>
<td>.0319</td>
<td>.0685</td>
<td>.0633</td>
</tr>
<tr>
<td>concen</td>
<td>.0306</td>
<td>.0579</td>
<td>.0579</td>
<td>.0486</td>
<td>.1519</td>
</tr>
<tr>
<td></td>
<td>.0097</td>
<td>.0534</td>
<td>.0534</td>
<td>.0118</td>
<td>.0578</td>
</tr>
</tbody>
</table>

- **POLS** - Pooled OLS, ignores route fixed effects (s.e. clustered by route)
- **FE** - FE estimator (s.e. clustered by route)
- **FE-BC** - Hahn-Kuersteiner bias-corrected estimator (s.e. clustered by route)
- **FD-IV** - FD estimator with $\log(\text{airfare}_{it-2})$ as instrument (s.e. clustered by route)
- **FD-ABIV** - Arellano-Bond GMM estimator, $\log(\text{airfare}_{it-2})$ as instrument in 1999, $\log(\text{airfare}_{it-2})$ and $\log(\text{airfare}_{it-3})$ as instruments in 2000, (heteroskedasticity robust s.e.)