Difference-in-Difference Models

Bill Evans
Health Economics

Florida

- 8/25/1997, State of Florida settles out of court in their suits against tobacco manufacturers
- Awarded $13 billion over 25 years
- Use $200m to run anti-smoking campaign aimed at kids
- Florida Tobacco Pilot Program (FTPP)
- Precursor to the national ‘truth’ campaign

- Florida’s edgy “Truth” advertising campaign continues to have a significant impact in reducing teen smoking, a team of researchers concluded from a new study that examines the impact of the state’s anti-tobacco advertising.

- in 1998, when surveillance began for tobacco use among Florida youth, 27.4 percent of high school students were current cigarette smokers. By 2000, this rate had declined to 22.6 among high school students.

- 4.8 percentage point decline or a 17.5% reduction in teen smoking
Nationwide

- Teen smoking rates fell from 36.5 to 31.4%
- A 5.1 percentage point decline or roughly 14%
- Rates in Florida fell by 4.8 percentage points – similar to what was happening in the nation as a whole
Difference in difference models

- Maybe the most popular “identification strategy” in applied statistical work in economics
- Attempts to mimic random assignment with treatment and “comparison” sample

Simple problem set up

- One group is ‘treated’ with intervention
- Have pre & post treatment data for group receiving intervention
- Can examine time-series changes but,
  - Unsure how much of the change is due to secular changes

- If the outcome of interest is trending over time, before/after comparisons will provide a biased estimate of the law
- Look at this graphically
Intervention occurs at time period $t_1$

- True effect of law
  - $Y_b - Y_a$

- Only have data at $t_1$ and $t_2$
  - If using time series, estimate of the effectiveness of the law is $Y_{t1} - Y_{t2}$

Solution?

**Difference in difference models**

- Pool cross-sectional and time series data
- Use time series of “untreated” group to establish “trends”
- What would have occurred in the treatment states in the absence of the intervention?

---

**Difference in Difference**

<table>
<thead>
<tr>
<th>Difference in Difference</th>
<th>Before Change</th>
<th>After Change</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>$Y_{t1}$</td>
<td>$Y_{t2}$</td>
<td>$\Delta Y_t = Y_{t2} - Y_{t1}$</td>
</tr>
<tr>
<td>Group 2</td>
<td>$Y_{c1}$</td>
<td>$Y_{c2}$</td>
<td>$\Delta Y_c = Y_{c2} - Y_{c1}$</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
<td>$\Delta \Delta Y = \Delta Y_t - \Delta Y_c$</td>
</tr>
</tbody>
</table>

**Motor Voter Example**

- Data in two years
  - 1992 Presidential (before MV)
  - 1996 Presidential (after)
- Two groups of states
  - Treated group (states that got MV through federal law in 1993)
  - Control group (states that had MV laws already)
Key Assumption

- Control group identifies the time path of outcomes that would have happened in the absence of the treatment.
- In this example, Y falls by $Y_{c2} - Y_{c1}$ even without the intervention.
- Note that underlying ‘levels’ of outcomes are not important (return to this in the regression equation).

In contrast, what is key is that the time trends in the absence of the intervention are the same in both groups.
- If the intervention occurs in an area with a different trend, will under/over state the treatment effect.
- In this example, suppose intervention occurs in area with faster falling Y.
Basic Econometric Model

- Data varies by
  - state (i)
  - time (t)
  - Outcome is Yt
- Only two periods
- Intervention will occur in a group of observations (e.g. states, firms, etc.)

Yt2 = \( \beta_0 + \beta_1 T_{it} + \beta_2 A_{it} + \beta_3 T_{it}A_{it} + \varepsilon_{it} \)

<table>
<thead>
<tr>
<th>Before Change</th>
<th>After Change</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1 (Treat)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2 (Control)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Meyer et al.

- Workers’ compensation
  - State run insurance program
  - Compensate workers for medical expenses and lost work due to on the job accident
- Premiums
  - Paid by firms
  - Function of previous claims and wages paid
- Benefits -- % of income w/ cap

Typical benefits schedule
- Min(pY,C)
- p=percent replacement
- Y = earnings
- C = cap
- e.g., 65% of earnings up to $400/week
- So if you earn >$616, only get $400/week

Concern:
- Moral hazard. Benefits will discourage return to work

Empirical question: duration/benefits gradient

Previous estimates
- Yi = β0 +Xβ1 +Rβ2 + εi
  - Y (duration)
  - R (replacement rate)
  - X (represents some other controls)
- Expect β2 > 0

Problem: Does realization of εi convey any information about R?
- Higher income workers, when injured, have much longer durations (εi>0)
- They also have lower replacement rates (Ri<average)
- Cov(Ri, εi)<0
- β2 is biased down
Solution

- Quasi experiment in KY and MI
- Increased the earnings cap
  - Increased benefit for high-wage workers
    - (Treatment)
  - Did nothing to those already below original cap
    (comparison)
- Compare change in duration of spell before and after change for these two groups

Data from Meyer et al.

- Data set kentucky.dta
- Key variables
  - durat (duration)
  - highearn (a high earning worker (treatment))
  - afchnge (after the law change)

* generate log duration
  . gen ldurat=ln(durat)

* sort the data by highearn and afchnge
  . sort highearn afchnge

* gets means of ldurat for
  . * 2x2 table
  . by highearn afchnge: sum ldurat
**Model**

- $Y_{it} = \text{duration of spell on WC}$
- $A_{it} = \text{period after benefits hike}$
- $H_{i} = \text{treated or high earnings group (Income}> E_3)$
- $Y_{it} = \beta_0 + H_{i}\beta_1 + A_{it}\beta_2 + A_{it}H_{i}\beta_3 + \varepsilon_{it}$
- Diff-in-diff estimate is $\beta_3$

### Difference in Difference

#### Mean average ln(duration)

<table>
<thead>
<tr>
<th></th>
<th>Before change</th>
<th>After change</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High earn</td>
<td>1.356</td>
<td>1.599</td>
<td>0.243</td>
</tr>
<tr>
<td>Low earn</td>
<td>1.123</td>
<td>1.137</td>
<td>0.014</td>
</tr>
<tr>
<td>Difference</td>
<td>0.233</td>
<td>0.462</td>
<td>0.229</td>
</tr>
</tbody>
</table>

### Model

- $Y_{it} = \beta_0 + H_{i}\beta_1 + A_{it}\beta_2 + A_{it}H_{i}\beta_3 + \varepsilon_{it}$
- Diff-in-diff estimate is $\beta_3$

### Code

- `* get the treatment affect`  
  `gen treat=highearn*afchnge`  
- `* run difference in difference regression`  
  `reg ldurat highearn afchnge treat`  

### Output

- Very low $R^2$
- Interpreting coefficients
- Compare results to 2 x 2 table – exactly the same
Making the model more complicated

- So far, a very simple model
  - Two groups
  - Two periods
- However, the “treatment” may cover more than 1 group
- The treatment may happen at very different time periods across groups
- How to generalize this type of model for
  - Many treatments
  - Multiple groups being treated

Example: States as laboratories

- Tremendous variation across states in their laws
  - Variation across states in any given year
  - Variation over time within a state
- Examples
  - Minimum wages, welfare policy, Medicaid coverage, traffic safety laws, use of death penalty, drinking age, cigarette taxes,

Empirical example:
Motorcycle Helmet laws

- 1967, Feds require states to have helmet law to get all federal highway money
- By 1975, all states have qualifying law
- 1976, Congress responds to state pressure and eliminate penalties
  - 20 states weaken their law and only require coverage for teens
  - 8 states repeal law completely

- 1991 Federal law again provides incentives for laws covering everyone
  - A bunch of states pass universal laws
- Congress changes its mind and in 1995 eliminate penalties
  - Again many states drop the law
- Currently
  - 20 states have universal law
  - 27 have teen coverage only
• Helmets are estimated to reduce the likelihood of death in a motorcycle crash by 37%. (Center for Disease Control)


• Where does this number come from?

FARS

• Fatal Accident Reporting System
• Census of motor vehicle accidents that produce a fatality
• Produced since 1975
• Detail information about
  – Accident
  – Vehicles
  – Drivers

• Select sample from FARS, 1988-2005
• Unique sample
  – Two riders on motor cycle
  – At least one died (accident was severe enough to produce a death)
  – Where one of the riders used a helmet, the other did not

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Key</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>frequency</td>
</tr>
<tr>
<td>row percentage</td>
</tr>
<tr>
<td>column percentage</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fatal</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>helmet</td>
<td>244</td>
<td>511</td>
<td>755</td>
</tr>
<tr>
<td></td>
<td>31.96</td>
<td>68.04</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>36.31</td>
<td>61.64</td>
<td>50.40</td>
</tr>
<tr>
<td></td>
<td>421</td>
<td>318</td>
<td>739</td>
</tr>
<tr>
<td></td>
<td>56.97</td>
<td>43.03</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>63.69</td>
<td>38.36</td>
<td>49.60</td>
</tr>
<tr>
<td>Total</td>
<td>661</td>
<td>829</td>
<td>1,490</td>
</tr>
<tr>
<td></td>
<td>44.36</td>
<td>55.64</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Pr(Die | no helmet) = 0.68
Pr(Die | helmet) = 0.43

Benefits of a helmet, reduce prob. of death by 37%

\[
\frac{0.43 - 0.68}{0.68} = -0.37
\]
Sources of bias using one-dimensional data

- Why would looking at one state over time provide a potentially biased estimate?

- Why might looking at the difference between states at a point in time provide a biased estimate?
Results for CA alone

```
reg mcdrl speed65 unemp bac_10 trend helmet_law if state=="CA"
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>Number of obs = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2.14855814</td>
<td>5</td>
<td>0.429711628</td>
<td>F(  5,    12) = 22.84</td>
</tr>
<tr>
<td>Residual</td>
<td>0.225732886</td>
<td>12</td>
<td>0.018811074</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>2.37429102</td>
<td>17</td>
<td>0.139664078</td>
<td>R-squared = 0.9049</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>Number of obs = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2.14855814</td>
<td>5</td>
<td>0.429711628</td>
<td>F(  5,    12) = 22.84</td>
</tr>
<tr>
<td>Residual</td>
<td>0.225732886</td>
<td>12</td>
<td>0.018811074</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>2.37429102</td>
<td>17</td>
<td>0.139664078</td>
<td>R-squared = 0.9049</td>
</tr>
</tbody>
</table>

| mcdrl | Coef. | Std. Err. | t  | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|----|-----|-----------------|
| speed65 | 0.4358863 | 0.1725389 | 2.53 | 0.027 | 0.0599564, 0.8118163 |
| unemp   | 0.0403951 | 0.0485866 | 0.70 | 0.497 | -0.0718221, 0.1399003 |
| bac_10  | 0.3175587 | 0.157151  | 2.02 | 0.066 | -0.0248439, 0.6599612 |
| trend   | 0.065022  | 0.0149018 | 4.36 | 0.001 | 0.0325538, 0.0974802 |
| helmet_law | -0.8972242 | 0.148954 | 6.02 | 0.000 | -1.264543, -0.5298851 |
| _cons   | -0.3775448 | 0.2939152 | 1.28 | 0.223 | -1.017931, 0.2628414 |

Motor cycle fatality rate -- treatment state

Motor cycle fatality rate -- comparison state

TX: Adopts in 1989, Repeals in 1998

CA: Adopts Helmet Law in 1992
Results for TX alone

* run a model for Texas
 . reg mcdrl speed65 unemp bac_08 trend helmet_law if state=="TX"

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>1.86115335</td>
<td>5</td>
<td>.37223067</td>
<td>Prob &gt; F = 0.0008</td>
</tr>
<tr>
<td>residual</td>
<td>.473677129</td>
<td>12</td>
<td>.039473094</td>
<td>R-squared = 0.7971</td>
</tr>
<tr>
<td>total</td>
<td>2.33483048</td>
<td>17</td>
<td>.137342969</td>
<td>Root MSE = .19868</td>
</tr>
</tbody>
</table>

------------------------------------------------------------------------------
<table>
<thead>
<tr>
<th></th>
<th>Coef.   Std. Err.      t    P&gt;</th>
<th>t</th>
<th></th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed65</td>
<td>.1689669   .2624034     0.64   0.532</td>
<td>-.4027609    .7406947</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unemp</td>
<td>.1301635   .0747096     1.74   0.107</td>
<td>-.0326147    .2929418</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bac_08</td>
<td>.6692796   .2457196     2.72   0.018</td>
<td>.1339026    1.204657</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trend</td>
<td>-.0405022   .0284775    -1.42   0.180</td>
<td>-.1025494     .021545</td>
<td></td>
<td></td>
</tr>
<tr>
<td>helmet_law</td>
<td>-.4592142   .1645968    -2.79   0.016</td>
<td>-.8178398   -.1005887</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>-.5765997    .464616    -1.24   0.238</td>
<td>-.1588911    .4357115</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

State group
Motor cycles registered per 100,000

<table>
<thead>
<tr>
<th>State group</th>
<th>Motor cycles per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never had helmet law</td>
<td>2366</td>
</tr>
<tr>
<td>Changed helmet law</td>
<td>1525</td>
</tr>
<tr>
<td>Always had helmet law</td>
<td>1224</td>
</tr>
</tbody>
</table>

Purely Cross Sectional Model (1990)

* run basic OLS model for 1990
 . reg mcdrl speed65 unemp bac_08bac_10 helmet_law if year==1990

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 48</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>2.59400681</td>
<td>5</td>
<td>.518801363</td>
<td>Prob &gt; F = 0.0023</td>
</tr>
<tr>
<td>residual</td>
<td>4.86098775</td>
<td>42</td>
<td>.115737804</td>
<td>R-squared = 0.3480</td>
</tr>
<tr>
<td>total</td>
<td>7.45499457</td>
<td>47</td>
<td>.158616906</td>
<td>Root MSE = .3402</td>
</tr>
</tbody>
</table>

------------------------------------------------------------------------------
<table>
<thead>
<tr>
<th></th>
<th>Coef.   Std. Err.      t    P&gt;</th>
<th>t</th>
<th></th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed65</td>
<td>.1692481   .1357937     1.25   0.220</td>
<td>-.1047947    .4432909</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unemp</td>
<td>.0524134   .0490639     1.07   0.292</td>
<td>-.0466015    .1514283</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bac_08</td>
<td>-.0944842   .2372792    -0.40   0.693</td>
<td>-.5733333   -.3843645</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bac_10</td>
<td>.01337 94.1658892     0.08   0.936</td>
<td>-.3213985   .3481573</td>
<td></td>
<td></td>
</tr>
<tr>
<td>helmet_law</td>
<td>-.4684841   .1038582    -4.51   0.000</td>
<td>-.6780784   -.2588898</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>-.0643492   .3042114    -0.21   0.833</td>
<td>-.6782726    .5405743</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

define:

\[
i = 1, 2, \ldots, n; \quad t = 1, 2, \ldots, T
\]

\[
S_i = 1 \text{ if state } i, = 0 \text{ otherwise}
\]

\[
W_t = 1 \text{ if year } t, = 0 \text{ otherwise}
\]

\[
\text{Law}_i = 1 \text{ if state } i \text{ has helmet law}
\]

\[
\text{in year } t, = 0 \text{ otherwise}
\]

\[
y_{it} = \beta_0 + \text{REFORM}_i \beta_1 + x_i \beta_2 + \sum_{j=2}^{n} \sum_{k=2}^{T} W_j \lambda_k + \epsilon_{it}
\]
• Why \( k = 2 \) to \( N \) and \( j = 2 \) to \( T \)?

• What does \( \alpha \) measure?

• What does \( \lambda \) measure?

• Question: impact of MC helmet laws on motorcycle fatalities

• Data: 48 states, 18 years (1988-2005), 864 observations

• Outcome \( \ln(\text{motorcycle death rate}) \)
  
  – Death rates = deaths/100,000 population

• Treatment variable: \( = 1 \) if state \( i \) has a motorcycle law in year \( t \), \( = 0 \) otherwise

Contains data from motorcycles.dta

| obs:  | 864 |
| size: | 49,248 (99.6% of memory free) |

<table>
<thead>
<tr>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>int</td>
<td>%9.0g</td>
<td>year</td>
<td></td>
</tr>
<tr>
<td>mcfatals</td>
<td>double</td>
<td>%9.0g</td>
<td>total motorcycle fatalities</td>
<td></td>
</tr>
<tr>
<td>state</td>
<td>str2</td>
<td>%2s</td>
<td>2 digit postal code, AL, CA, etc.</td>
<td></td>
</tr>
<tr>
<td>fips</td>
<td>byte</td>
<td>%8.0g</td>
<td>2 digit numeric fips code</td>
<td></td>
</tr>
<tr>
<td>helmet_law</td>
<td>float</td>
<td>%9.0g</td>
<td>=1 if motorcycle helmet law, ( =0 ) otherwise</td>
<td>helmet_law</td>
</tr>
<tr>
<td>speed65</td>
<td>float</td>
<td>%9.0g</td>
<td>=1 if speed limit is 65, ( =0 ) otherwise</td>
<td>speed65</td>
</tr>
<tr>
<td>speed70p</td>
<td>float</td>
<td>%9.0g</td>
<td>=1 if speed limit is 70 plus, ( =0 ) otherwise</td>
<td>speed70p</td>
</tr>
<tr>
<td>bac_10</td>
<td>float</td>
<td>%9.0g</td>
<td>drunk driving defined as bac=0.1, ( =0 ) otherwise</td>
<td>bac_10</td>
</tr>
<tr>
<td>bac_08</td>
<td>float</td>
<td>%9.0g</td>
<td>drunk driving defined as bac=0.08, ( =0 ) otherwise</td>
<td>bac_08</td>
</tr>
<tr>
<td>unemp</td>
<td>float</td>
<td>%9.0g</td>
<td>state unemployment rate, 5 is 5%</td>
<td></td>
</tr>
<tr>
<td>population</td>
<td>float</td>
<td>%9.0g</td>
<td>state population</td>
<td></td>
</tr>
<tr>
<td>mregs</td>
<td>double</td>
<td>%10.0g</td>
<td>motorcycle registrations</td>
<td></td>
</tr>
</tbody>
</table>

Sorted by:

\[ \text{state} \times \text{year} \]

\[ \_Istate_1-48 \ (\_Istate_1 \text{ for state=AL omitted}) \]

\[ \_Iyear_1988-2005 \ (\text{naturally coded}; \_Iyear_1988 omitted) \]

\[ \text{state} \times \text{year} \]

\[ \_Istate_1-48 \ (\_Istate_1 \text{ for state=AL omitted}) \]

\[ \_Iyear_1988-2005 \ (\text{naturally coded}; \_Iyear_1988 omitted) \]

\[ \text{state} \times \text{year} \]

\[ \_Istate_1-48 \ (\_Istate_1 \text{ for state=AL omitted}) \]

\[ \_Iyear_1988-2005 \ (\text{naturally coded}; \_Iyear_1988 omitted) \]
```
. * run the difference in difference model
. reg mcdrl speed65 speed70p unemp bac_08 bac_10 _I* helmet_law

Source |       SS       df       MS              Number of obs =   864
--------+------------------------------ F( 70,   793) =  30.54
Model |  139.812929    70  1.99732756           Prob > F      =  0.0000
Residual |  51.8558902   793  .065392043           R-squared     =  0.7295
--------+------------------------------ Adj R-squared =  0.7056
Total |   191.66882   863  .222095967           Root MSE      =  .25572

------------------------------------------------------------------------------
mcdrl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
speed65 |  -.0577686   .0552537    -1.05   0.296    -.1662293    .0506922
speed70p |  -.0855586   .0815308    -1.05   0.294    -.2456004    .0744831
unemp |  -.0117339   .0118625    -0.99   0.323    -.0350195    .0115517
bac_08 |   .1423512   .0725064     1.96   0.050     .0000241    .2846783
bac_10 |   .1163134   .0628129     1.85   0.064    -.0069859    .2396127
_Istate_2 |  -.038139   .0889074    -0.43   0.668    -.2126606    .1363826
delete some results
_Istate_47 |   .2392712   .0896769     2.67   0.008     .0632391    .4153033
_Istate_48 |   .3987819   .0978800     4.07   0.000     .2066474    .5909164
_Iyear_1989 |  -.2367341    .052373    -4.52   0.000    -.3395401   -.1339281
delete some results
_Iyear_2005 |   .1032509   .0703676     1.47   0.143    -.0348778    .2413796
helmet_law |  -.3728078   .0458932    -8.12   0.000    -.4628943   -.2827213
delete some results
_cons |   .5393718   .1275965     4.23   0.000     .2889049    .7898387
------------------------------------------------------------------------------
```