**Past & Future**

**Last Class**
- Panel Data: intuition
- Fixed Effects: FD
- Fixed Effects: Demeaning
- Random Effects

**Today**
- DiD: two by two
- DiD: general case
- Synthetic Controls
- Discussion
Today we will focus again on the problem of evaluating the impact of a program or treatment on a population outcome $Y$.

**Potential outcomes:**

$Y(0)$  potential outcome in the absence of treatment

$Y(1)$  potential outcome in the presence of treatment

The treatment effect is the difference $Y(1) - Y(0)$ and a popular quantity of interest is $E[Y(1) - Y(0)]$, typically referred to as the average treatment effect.

A large fraction of the work in econometric theory precisely deals with deriving methods that may recover the average treatment affect (or similar quantities) from observing $Y_i(1)$ for individuals receiving treatment and $Y_i(0)$ for individuals without treatment (but never both).

The difference in differences (DD) approach is a popular method in this class that exploits grouped-level treatment assignments that vary over time.
A Simple Two by Two Case

- **2x2**: simplest setup to describe the DD approach is one where outcomes are observed for two groups for two time periods.

- **Group 1 treated**: group 1 is exposed to a treatment in the second period but not in the first period.

- **Group 2 untreated**: group 2 is not exposed to the treatment during either period.

\[(Y_{j,t}, D_{j,t}) : j \in \{1, 2\} \text{ and } t \in \{1, 2\}\]

is the observed data, where \(Y_{j,t}\) and \(D_{j,t}\) denote the outcome and treatment of \(j\) at time \(t\).

- **Treatment**: \(D_{j,t} = I[j = 1, t = 2]\)

- The parameter we will be able to identify is

\[\theta = E[Y_{1,2}(1) - Y_{1,2}(0)],\]

which is simply the **average treatment effect on the treated** (group 1 in period 2).
On April 1, 1992, New Jersey raised the state minimum wage from $4.25 to $5.05.

Card and Krueger (1994) collected data on employment at fast food restaurants in New Jersey in February 1992 \((t = 1)\) and again in November 1992 \((t = 2)\) to study the effect of increasing the minimum wage on employment.

They also collected data from the same type of restaurants in eastern Pennsylvania, just across the river. The minimum wage in Pennsylvania stayed at $4.25 throughout this period.

In our notation, New Jersey would be the first group, \(Y_{j,t}\) would be the employment rate in group \(j\) at time \(t\), and \(D_{j,t}\) denotes an increase in the minimum wage (the treatment) in group \(j\) at time \(t\).
The identification strategy of DD relies on the **Common Trends** assumption

\[ E[Y_{2,2}(0) - Y_{2,1}(0)] = E[Y_{1,2}(0) - Y_{1,1}(0)] . \]

Both groups have “common trends” in the absence of a treatment.

One way to parametrize this assumption is

\[ Y_{j,t}(0) = \eta_j + \gamma_t + U_{j,t} , \]

where \( E[U_{j,t}] = 0 \), and \( \eta_j \) and \( \gamma_t \) are (non-random) group and time effects.

**Note:** \( E[Y_{j,2}(0) - Y_{j,1}(0)] = \gamma_2 - \gamma_1 \equiv \gamma \), which is constant across groups. In addition,

\[ E[Y_{1,2}(1)] = \theta + \eta_1 + \gamma_2 . \]

In the example: in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect, a year effect that is common across states, and a zero mean shock.
$Y_{j,t}(0) = \eta_j + \gamma_t + U_{j,t}$ where $E[Y_{1,2}(1)] = \theta + \eta_1 + \gamma_2$. 

**Pre and post comparison**
Treatment and control comparison

\[ Y_{j,t}(0) = \eta_j + \gamma_t + U_{j,t} \quad \text{where} \quad E[Y_{1,2}(1)] = \theta + \eta_1 + \gamma_2. \]
Taking both differences

\[ Y_{j,t}(0) = \eta_j + \gamma_t + U_{j,t} \quad \text{where} \quad E[Y_{1,2}(1)] = \theta + \eta_1 + \gamma_2. \]
Causal effects in the DD model

\[ t = 1 \]

\[ t = 2 \]

**Figure**: Causal effects in the DD model
QUESTIONS?
Suppose that we observe

$$\{(Y_{i,j,t}, D_{j,t}) : i \in I_{j,t}, j \in \{1, 2\} \text{ and } t \in \{1, 2\}\},$$

where $I_{j,t}$ is the set of individuals in group $j$ at time $t$.

Take $D_{j,t} = I\{j = 1\}I\{t = 2\}$ to be non-random and note that the observed outcome is

$$Y_{i,j,t} = Y_{i,j,t}(1)D_{j,t} + (1 - D_{j,t})Y_{i,j,t}(0) = (Y_{i,j,t}(1) - Y_{i,j,t}(0))D_{j,t} + Y_{i,j,t}(0),$$

so that if we define $U_{i,j,t} = Y_{i,j,t} - E[Y_{i,j,t}]$, we can write

$$Y_{i,j,t} = \theta D_{j,t} + \eta_j + \gamma_t + U_{i,j,t}$$

We can estimate $\theta$ by running a regression of $Y_{i,j,t}$ on $D_{j,t}$ that includes units and time fixed effects.

The regression formulation of the DD model offers a convenient way to construct DD estimates and standard errors. It also makes it easy to add additional groups and time periods.
Many groups, many time periods (and no individual data for now). The analog regression is

\[ Y_{j,t} = \theta D_{j,t} + \eta_j + \gamma_t + U_{j,t} \quad \text{with} \quad E[U_{j,t}] = 0. \]

**Observed data:** \{\((Y_{j,t}, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\)\} where

- \(T_0\) is the set of pre-treatment time periods,
- \(T_1\) is the set of post-treatment time periods,
- \(J_0\) is the set of controls units,
- \(J_1\) is the set of treatment units.

The random variables \(\eta_j, \gamma_t\) and \(U_{j,t}\) are unobserved and \(\theta \in \Theta \subseteq \mathbb{R}\) is the parameter of interest.

Define

\[ \Delta_{n,j} = \frac{1}{|J_1|} \sum_{t \in J_1} Y_{j,t} - \frac{1}{|J_0|} \sum_{t \in J_0} Y_{j,t}, \]

and

\[ \hat{\theta}_n = \frac{1}{|J_1|} \sum_{j \in J_1} \Delta_{n,j} - \frac{1}{|J_0|} \sum_{j \in J_0} \Delta_{n,j}. \]
A More General Case

- **LS**: Easy to show that $\hat{\theta}_n$ is the LS estimator of a regression of $Y_{j,t}$ on $D_{j,t}$ with groups fixed effects ($\eta_j$) and time fixed effects ($\gamma_t$).

- **By simple algebra**:

  $$\hat{\theta}_n - \theta = \frac{1}{|J_1|} \sum_{j \in J_1} \left( \frac{1}{|T_1|} \sum_{t \in T_1} U_{j,t} - \frac{1}{|T_0|} \sum_{t \in T_0} U_{j,t} \right) - \frac{1}{|J_0|} \sum_{j \in J_0} \left( \frac{1}{|T_1|} \sum_{t \in T_1} U_{j,t} - \frac{1}{|T_0|} \sum_{t \in T_0} U_{j,t} \right).$$

- It follows immediately from $E[U_{j,t}] = 0$ that $E[\hat{\theta}_n] = \theta$.

- This estimator is also **consistent** and **asymptotically normal** in an asymptotic framework with a large number of treated and untreated groups, i.e., $|J_1| \to \infty$ and $|J_0| \to \infty$.

- The parameter $\theta$ could be interpreted as the ATT under the assumption that

  $$E[Y_{j,t}(1) - Y_{j,t}(0)],$$

  is **constant** for all $j \in J_1$ and $t \in T_1$. Alternatively, one could estimate a different $\theta_j$ for each $j \in J_1$. 

Inference in DD could be tricky and requires thinking. **Two issues** are of particular importance.

1. **First**: what exactly is assumed to be “large”? Are groups going to infinity? Say, $|J_1| \to \infty$ and $|J_0| \to \infty$.

   What happens if we have a few treated groups but many controls? Say, $|J_1|$ fixed and $|J_0| \to \infty$.

   What happens if we have few treated and control groups but many time periods? Say, $|J_1|$ and $|J_0|$ fixed, but $|T_1| \to \infty$ and $|T_0| \to \infty$.

2. **Second**: Time dependence. It is typically common to assume that $U_{j,t} \perp U_{j',s}$ for all $j' \neq j$ and $(t,s)$.

   However, one would expect $U_{j,t}$ and $U_{j,s}$ to be correlated, at least for $t$ and $s$ being “close” to each other.

   Also, with individual data one would expect $U_{i,j,t}$ to be correlated with $U_{i',j,s}$ - i.e., units in the same group may be dependent to each other even if they are in different time periods.

Each of these aspects have tremendous impact on which inference tools end up being valid or not.
Illustration: consider $J_1 = \{1\}$ and $|J_0| \to \infty$ - also assume $|\mathcal{T}_0|$ and $|\mathcal{T}_1|$ are finite.

The DD estimator in this case reduced to

$$\hat{\theta}_n = \Delta_{n,1} - \frac{1}{|J_0|} \sum_{j \in J_0} \Delta_{n,j},$$

$$= \theta + \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} U_{1,t} - \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} U_{1,t} - \frac{1}{|J_0|} \sum_{j \in J_0} \left( \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} U_{j,t} - \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} U_{j,t} \right)$$

$$\xrightarrow{p} \theta + \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} U_{1,t} - \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} U_{1,t},$$

as $|J_0| \to 0$, assuming $\{U_{j,t} : t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$ is i.i.d. across $j \in J_0$.

Result: the DD estimator is not even consistent for $\theta$. Still possible to do inference on $\theta$ using the approach proposed by Conley and Taber (2011) or, more recently, the randomization approach in Canay, Romano, and Shaikh (2017).
QUESTIONS?
On the common trends assumption

- Keep in mind that all the results on DD follow from the assumption that

\[ E[Y_{j,t}(0)] = \eta_j + \gamma_t , \]

which is a way to model the “common trends” assumption.

- Where there are multiple time periods, people will often look at the pre (and post) treatment trends and compare them between treatment and control as a way to “eye-ball” verify this assumption.

- **Unpleasant feature**: it is not robust to nonlinear transformations of the outcome variables, i.e.,

\[ E[Y_{2,2}(0) - Y_{2,1}(0)] = E[Y_{1,2}(0) - Y_{1,1}(0)] , \]

does not imply, for example, that

\[ E[\log Y_{2,2}(0) - \log Y_{2,1}(0)] = E[\log Y_{1,2}(0) - \log Y_{1,1}(0)] . \]

- These are non-nested and one would typically suspect that both cannot hold at the same time.
**Synthetic Controls**

- **DD approach**: (i) treats all control groups as being of equal quality as a control group, (ii) requires common trends.

**Synthetic Controls**

- The researcher may want to somehow weight the controls in order to give more importance to those controls that seem “better” for the given treated group.

- Allows for a model with interactive effects (no common trends):
  
  \[ Y_{j,t} = \eta_j \gamma_t + U_{j,t} . \]

- Originally proposed by Abaide, et al (2010, ADH) to study the effect of California’s tobacco control program on state-wide smoking rates. During the time period in question, there were 38 states in the US that did not implement such programs.

- ADH propose choosing a weighted average of the potential controls, formalizing a procedure that optimally chooses weights.
Synthetic Controls

- **2x2 model**: except that now there are $J_0 > 2$ possible controls and that $J_1 = \{1\}$.

- **Naive comparison**: comparing $Y_{1,2}$ and $Y_{j,2}$ for any $j \in J_0$ delivers

  $$E[Y_{1,2} - Y_{j,2}] = E[Y_{1,2}(1) - Y_{j,2}(0)] = \theta + \gamma_2(\eta_1 - \eta_j),$$

  and so this approach does not identify $\theta$ in the presence of persistent group differences.

- The idea behind **synthetic controls** is to construct the so-called *synthetic control*

  $$\tilde{Y}_{1,2}(0) = \sum_{j \in J_0} w_j Y_{j,2},$$

  by appropriately choosing the **weights** $\{w_j : j \in J_0, w_j \geq 0, \sum_{j \in J_0} w_j = 1\}$.

- In order for this idea to work, it must be the case that

  $$E[Y_{1,2}(0)] = E[\tilde{Y}_{1,2}(0)] \Rightarrow E[Y_{1,2} - \tilde{Y}_{1,2}(0)] = \theta.$$
Let \( \{w_j : j \in J_0, w_j \geq 0, \sum_{j \in J_0} w_j = 1\} \) be given. This approach delivers

\[
E \left[ Y_{1,2} - \bar{Y}_{1,2}(0) \right] = E \left[ Y_{1,2} - \sum_{j \in J_0} w_j Y_{j,2} \right] = \theta + \gamma_2 \left( \eta_1 - \sum_{j \in J_0} w_j \eta_j \right).
\]

The approach then identifies \( \theta \) if we could choose the weights in a way such that

\[
\eta_1 = \sum_{j \in J_0} w_j \eta_j.
\]

**Infeasible:** we do not observe the group effects \( \eta_j \).

**Result in ADH:** suppose that there exists weights \( \{w^*_j : j \in J_0, w^*_j \geq 0, \sum_{j \in J_0} w^*_j = 1\} \) such that

\[
Y_{1,1} = \sum_{j \in J_0} w^*_j Y_{j,1}.
\]

Then we can identify \( \theta \) by using these weights:

\[
\bar{Y}_{1,2}(0) = \sum_{j \in J_0} w^*_j Y_{j,2} \quad \Rightarrow \quad E \left[ Y_{1,2} - \bar{Y}_{1,2}(0) \right] = \theta.
\]
Proof in the Example

\[ Y_{1,1} = \sum_{j \in \mathcal{J}_0} w_j^* Y_{j,1} \quad \text{and} \quad \tilde{Y}_{1,2}(0) = \sum_{j \in \mathcal{J}_0} w_j^* Y_{j,2} \quad \text{lead to} \quad E[Y_{1,2} - \tilde{Y}_{1,2}(0)] = 0 \]
Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

FIGURE: Synthetic Controls in Action
**Weights**: by “matching” the observed outcomes of the treated group and the control groups in the periods before the policy change.

$Y_{1,1}$ may not lie in the convex hull of $\{Y_{j,1} : j \in J_0\}$. Method relies on minimizing the distance between $Y_{1,1}$ and $\sum_{j \in J_0} w_j Y_{j,1}$.

ADH provide formal arguments. They require that $|J_0| \to \infty$ and $U_{j,t}$ independent across $j$ and $t$.

**Important I**: the model they consider does not require the “common trends” assumption

**Important II**: formal arguments account for randomness of the weights.

**Covariates**: can be extended in the presence of covariates $X_j$ that are not (or would not be) affected by the policy change. In this case, the weights would be chosen to minimize the distance between

$$(Y_{1,1}, X_1) \text{ and } \sum_{j \in J_0} w_j (Y_{j,1}, X_j).$$

The optimal weights - which differ depending on how we define distance - produce the synthetic control whose pre-intervention outcome and predictors of post-intervention outcome are “closest”.
THE END!