ECMA 31100: Intro to Empirical Analysis II

Difference in Differences

Joe Hardwick

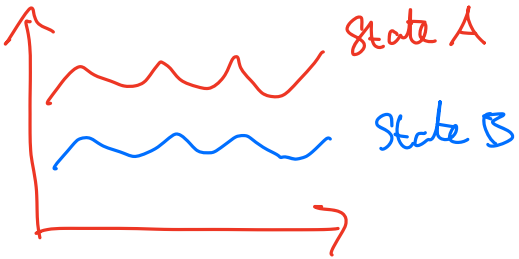
University of Chicago

Winter 2022

Introduction

- Previously considered potential outcomes $y(1)$ and $y(0)$, which denote what *would* occur in the presence or absence of treatment at a particular period in time.
- Not generally possible to estimate ATE using naive comparison, unless there is random assignment.
- May also estimate ATE if there is random assignment conditional on observables.
- Today we identify the ATT under a different assumption: Parallel trends in outcome over time.

Introduction



- Parallel trends allows for differences in levels of potential outcomes across groups, but assumes in the absence of treatment that trends would have been the same.
- Even if treatment is not randomly assigned, the *trends* are as good as randomly assigned
- May be plausible if, say, policy changes at state level, so states will differ in *level* of outcome, but may not differ in trends.

Data

- Comparing trends requires us to have a cross-section of outcome data before the treatment occurs, and a cross section after.
- Observing a random sample of observations in several time periods is called a repeated cross section.
- Observing the *same* units in several time periods is called panel data.
- In a repeated cross-section there is no need for the units to be the same over time.
    - Every panel is a repeated cross-section, but not vice versa

Example: Store Upgrades

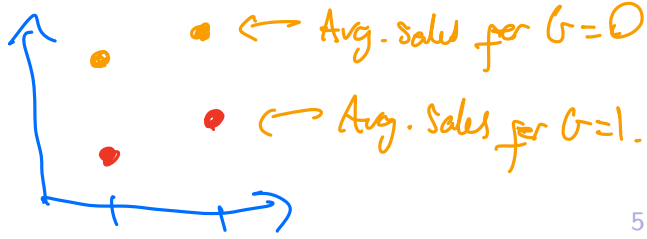
Groups $G = 0, 1$
Times $T = 0, 1$

- Suppose manager chooses to upgrade stores with low sales figures.
- There are 2 time periods, before and after.
- We observe sales figures for stores that were upgraded ($G_i = 1$) and stores that were not upgraded ($G_i = 0$), before and after upgrades:
- We have observations before upgrades ($T_i = 0$) and after upgrades ($T_i = 1$).
- Want

$$ATT = \mathrm{E}\left(y\left(1\right) - y\left(0\right) | G = 1, T = 1\right) :$$

Difference in sales for stores that were upgraded after they were upgraded

$E(y_1 - y_0 | D=1)$ $\quad D = G \cdot T$



Example: Store Upgrades

T=0 T=1

- Why not just compare sales for upgraded stores before and after?
  - Temporal variation in sales due to seasonality or other changes made at the same time upgrades were done
  - End up confounding e.g. seasonality with effect of treatment.
- Let Y be the observed outcome.

$$Y = \begin{cases} y(1) & \text{if } G = 1, T = 1, \\ y(0) & \text{otherwise.} \end{cases}$$

$Y = Dy_1 + (1-D)y_0$.

Example: Store Upgrades

- Comparison of means across time:

$$
\begin{aligned}
&\mathrm{E}(Y|G=1,T=1)-\mathrm{E}(Y|G=1,T=0)\\
=&\mathrm{E}(Y|G=1,T=1)-\mathrm{E}(y(0)|G=1,T=1)\\
+&\mathrm{E}(y(0)|G=1,T=1)-\mathrm{E}(Y|G=1,T=0)\\
=&\underbrace{\mathrm{E}(y(1)-y(0)|G=1,T=1)}_{ATT}\\
+&\underbrace{\mathrm{E}(y(0)|G=1,T=1)-\mathrm{E}(y(0)|G=1,T=0)}_{\text{Trend in demand for upgraded stores in absence of upgrade}}
\end{aligned}
$$

A naive comparison across time periods gives the ATT + temporal variation.

Example: Store Upgrades

- Why not just compare stores across treated and control, after the upgrade?
- Comparison of means across groups:

$$
\begin{aligned}
&\mathrm{E}(Y|G=1,T=1)-\mathrm{E}(Y|G=0,T=1)\\
=&\mathrm{E}(Y|G=1,T=1)-\mathrm{E}(y(0)|G=1,T=1)\\
+&\mathrm{E}(y(0)|G=1,T=1)-\mathrm{E}(Y|G=0,T=1)\\
=&\underbrace{\mathrm{E}(y(1)-y(0)|G=1,T=1)}_{ATT}\\
+&\underbrace{\mathrm{E}(y(0)|G=1,T=1)-\mathrm{E}(y(0)|G=0,T=1)}_{\text{Selection Bias in } y(0)}
\end{aligned}
$$

A naive comparison across groups gives the ATT + selection bias in $y(0)$.

Common Trends

- Common Trends Assumption: In the absence of treatment, the trend in sales for upgraded stores and non-upgraded stores would have been the same.
- Formally:

Unobservable

$$\underbrace{\mathrm{E}\left(y\left(0\right)|G=1,T=1\right)-\mathrm{E}\left(y\left(0\right)|G=1,T=0\right)}_{\text{Trend in demand for upgraded stores in absence of upgrade}}$$

$$=\underbrace{\mathrm{E}\left(y\left(0\right)|G=0,T=1\right)-\mathrm{E}\left(y\left(0\right)|G=0,T=0\right)}_{\text{Trend in demand for non-upgraded stores in absence of upgrade}}$$

$\beta_0+\beta_1+\beta_2+\beta_3-(\beta_0+\beta_1)$

$=\beta_0+\beta_2-\beta_0$

$\beta_2+\beta_3=\beta_2$

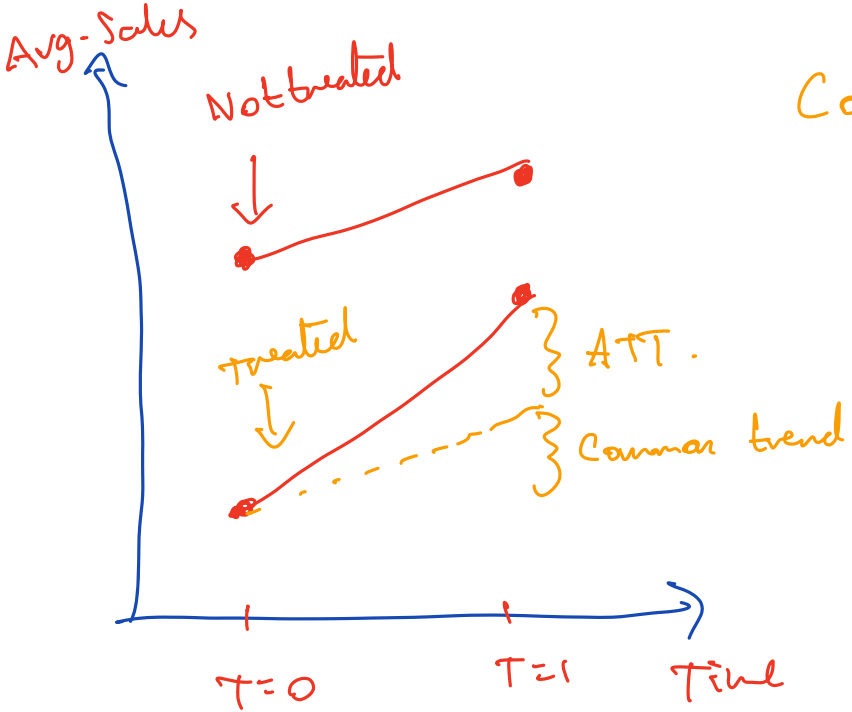
$\beta_3=0.$

Common Trends

- Under this assumption, naive comparison across time gives:

$$
\begin{aligned}
&\mathrm{E}\left(Y|G=1,T=1\right)-\mathrm{E}\left(Y|G=1,T=0\right)\\
=&\underbrace{\mathrm{E}\left(y\left(1\right)-y\left(0\right)|G=1,T=1\right)}_{ATT}\\
+&\underbrace{\mathrm{E}\left(y\left(0\right)|G=0,T=1\right)-\mathrm{E}\left(y\left(0\right)|G=0,T=0\right)}_{\text{Trend in demand for non-upgraded stores in absence of upgrade}}\\
&\underbrace{\mathrm{E}\left(y\left(0\right)|G=1,T=1\right)-\mathrm{E}\left(y\left(0\right)|G=1,T=0\right)}_{\text{Trend in demand for upgraded stores in absence of upgrade}}
\end{aligned}
$$

Example: Store Upgrades



Common trend can be estimated, so the ATT can also.

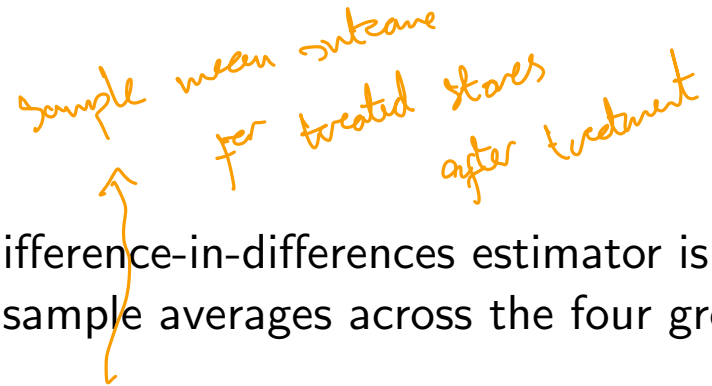
Identifying ATT

- Now compute difference in trends for treated and control groups:

$$
\begin{aligned}
\text{Diff-in-Diff} =& \left[\mathrm{E}\left(Y|G=1,T=1\right)-\mathrm{E}\left(Y|G=1,T=0\right)\right] \\
&-\left[\mathrm{E}\left(Y|G=0,T=1\right)-\mathrm{E}\left(Y|G=0,T=0\right)\right] \\
=& \left[\mathrm{E}\left(Y|G=1,T=1\right)-\mathrm{E}\left(Y|G=1,T=0\right)\right] \\
&-\left[\underbrace{\mathrm{E}\left(y\left(0\right)|G=0,T=1\right)-\mathrm{E}\left(y\left(0\right)|G=0,T=0\right)}_{\text{Trend in demand for non-upgraded stores in absence of upgrade}}\right] \\
=& ATT.
\end{aligned}
$$

Questions?

Regression



- The difference-in-differences estimator is implemented by using sample averages across the four groups:

$$\widehat{ATT} = \left[\bar{Y}_{G=1,T=1} - \bar{Y}_{G=1,T=0}\right] - \left[\bar{Y}_{G=0,T=1} - \bar{Y}_{G=0,T=0}\right].$$

  We can use linear regression to generate this estimate.

- First, note that $\mathrm{E}(y(0)\,|\,G, T)$ can take 4 values, so WLOG:

$$\mathrm{E}(y(0)\,|\,G, T) = \beta_0 + \beta_1 G + \beta_2 T + \beta_3 G \cdot T.$$

- The common trends assumption holds iff $\beta_3 = 0$.

Regression

$Y = y(0) + D(y(1) - y(0))$

- Now we write the conditional mean of the observed outcome:

$$\begin{aligned}\mathrm{E}(Y|G,T) &= \mathrm{E}(y(0)|G,T) + \mathrm{E}(y(1) - y(0)|G,T) \cdot G \cdot T \\ &= \mathrm{E}(y(0)|G,T) \\ &+ \mathrm{E}(y(1) - y(0)|G=1,T=1) \cdot G \cdot T \\ &= \mathrm{E}(y(0)|G,T) + ATT \cdot G \cdot T\end{aligned}$$

since $y(1)$ is only observed if $G = T = 1$. Hence:

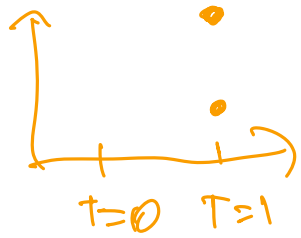
$$\mathrm{E}(Y|G,T) = \beta_0 + \beta_1 G + \beta_2 T + [\beta_3 + ATT]\, G \cdot T,$$

so under common trends, the coefficient on $G \cdot T$ is the ATT.

Effect of Incinerator on Home Prices

- Wooldridge Example 13.3: Have home values in 1978 ($y81 = 0$) and in 1981 ($y81 = 1$), when a new garbage incinerator was built.
- Hypothesise that house prices fell as a result of the incinerator.
- House defined to be near the incinerator (*nearinc*) if within 3 miles.
- Price (*rprice*) measured in 1978 dollars for each house.
- Can naively compare house prices in 1981:

Effect of Incinerator on Home Prices




```
. reg rprice nearinc if year==1981
```

Source	SS	df	MS		
Model	2.7059e+10	1	2.7059e+10	Number of obs	= 142
Residual	1.3661e+11	140	975815048	F(1, 140)	= 27.73
				Prob > F	= 0.0000
Total	1.6367e+11	141	1.1608e+09	R-squared	= 0.1653
				Adj R-squared	= 0.1594
				Root MSE	= 31238

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97	-19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43	107422.6

Difference in avg. prices after incinerator built ≈$31,000

Effect of Incinerator on Home Prices

- This does not imply the incinerator caused a reduction in house price of $30,688, merely that average house prices near the incinerator were $30,688 lower on average.
- Run the same regression for house prices in 1978, before the incinerator:

```
. reg rprice nearinc if year==1978
```

Source	SS	df	MS		
Model	1.3636e+10	1	1.3636e+10	Number of obs	= 179
Residual	1.5332e+11	177	866239953	F(1, 177)	= 15.74
				Prob > F	= 0.0001
				R-squared	= 0.0817
				Adj R-squared	= 0.0765
Total	1.6696e+11	178	937979126	Root MSE	= 29432

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-18824.37	4744.594	-3.97	0.000	-28187.62	-9461.117
_cons	82517.23	2653.79	31.09	0.000	77280.09	87754.37

Difference in average price of $\approx \$18,000$.

Effect of Incinerator on Home Prices

- These results show that homes near the incinerator had lower values on average, so the incinerator may have been built in an area with lower housing values, indicating selection bias.
- Diff-in-Diff compares the difference in house prices before and after the incinerator was built:
- Define

$$y81 = \begin{cases} 1 & \text{if } year = 1981 \\ 0 & \text{if } year = 1978 \end{cases}.$$

Can run a regression that allows for different intercepts and slopes for each housing area.

Effect of Incinerator on Home Prices

$E(y_0 | G, T) = \beta_0 + \beta_1 G + \beta_2 T + \beta_3 G \cdot T$.

- Model:

$$\begin{aligned} y &= \beta_0 + \beta_1 \cdot y81 + nearinc\,(\gamma_0 + \gamma_1 y81) + v \\ &= \beta_0 + \beta_1 \cdot y81 + \gamma_0 nearinc + \gamma_1 y81 \cdot nearinc + v. \end{aligned}$$

  Under the assumption of common trends, γ_1 is the ATT, the effect of the incinerator on house prices near where it was built.

Effect of Incinerator on Home Prices

inter = y81 · nearinc ·

```
. reg rprice y81 nearinc inter
```

Source	SS	df	MS	Number of obs	=	321
				F(3, 317)	=	22.25
Model	6.1055e+10	3	2.0352e+10	Prob > F	=	0.0000
Residual	2.8994e+11	317	914632739	R-squared	=	0.1739
				Adj R-squared	=	0.1661
Total	3.5099e+11	320	1.0969e+09	Root MSE	=	30243

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
nearinc	-18824.37	4875.322	-3.86	0.000	-28416.45	-9232.293
inter	-11863.9	7456.646	-1.59	0.113	-26534.67	2806.867
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

Effect on houses built near incinerator is estimated to be ≈ -$12,000.

Including Controls

- We may include controls in the main regression to:
  - Modify the common trends assumption, so it now must hold conditional on covariates
  - Reduce the standard error on the interaction.
- How do our assumptions change? Let G denote houses near incinerator, $T = 1$ in 1981.
  - Common trends conditional on controls X:

$$\underbrace{\mathrm{E}\left(y\left(0\right)|G=1,T=1,X=x\right)-\mathrm{E}\left(y\left(0\right)|G=1,T=0,X=x\right)}_{\text{Trend in price for houses near incinerator in absence of incinerator}}$$

$$=\underbrace{\mathrm{E}\left(y\left(0\right)|G=0,T=1,X=x\right)-\mathrm{E}\left(y\left(0\right)|G=0,T=0,X=x\right)}_{\text{Trend in price for houses far from incinerator in absence of incinerator}}$$

  - Doesn't imply common trends because covariates can adjust over time differently for both groups.

Including Controls

Incorporates common trends assumption conditional on X. (no β_3).

- e.g. Suppose

$$\mathrm{E}\left(y\left(0\right)|G,T,X\right)=\beta_0+\beta_1 G+\beta_2 T+X'\delta.$$

Then

$\beta_0+\beta_1+\beta_2+\mathrm{E}[X|G=1,T=1)'\delta$.

$\beta_0+\beta_1+\mathrm{E}[X|G=1,T=0)'\delta$

$$\begin{aligned}&\mathrm{E}\left(y\left(0\right)|G=1,T=1\right)-\mathrm{E}\left(y\left(0\right)|G=1,T=0\right)\\&=\beta_2+\left[\mathrm{E}\left(X|G=1,T=1\right)-\mathrm{E}\left(X|G=1,T=0\right)\right]'\delta\end{aligned}$$

The same calculation for $G=0$ shows that common trends only holds if the composition of covariates across time changes in the same way in both groups.

- Including covariates therefore allows us to identify the ATT if common trends conditional on X holds.

Including Controls

- If the composition of covariates doesn't change differently across groups, the standard error may still reduce significantly as a result of their inclusion:

Source	SS	df	MS		
				Number of obs	= 321
				F(10, 310)	= 60.19
Model	2.3167e+11	10	2.3167e+10	Prob > F	= 0.0000
Residual	1.1932e+11	310	384905860	R-squared	= 0.6600
				Adj R-squared	= 0.6491
Total	3.5099e+11	320	1.0969e+09	Root MSE	= 19619

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	13928.48	2798.747	4.98	0.000	8421.533	19435.42
nearinc	3780.337	4453.415	0.85	0.397	-4982.408	12543.08
inter	-14177.93	4987.267	-2.84	0.005	-23991.11	-4364.759
age	-739.451	131.1272	-5.64	0.000	-997.4629	-481.4391
agesq	3.45274	.8128214	4.25	0.000	1.853395	5.052084
intst	-.5386352	.1963359	-2.74	0.006	-.9249548	-.1523157
land	.1414196	.0310776	4.55	0.000	.0802698	.2025693
area	18.08621	2.306064	7.84	0.000	13.54869	22.62373
rooms	3304.227	1661.248	1.99	0.048	35.47904	6572.974
baths	6977.317	2581.321	2.70	0.007	1898.191	12056.44
_cons	13807.67	11166.59	1.24	0.217	-8164.239	35779.57

Questions?

Triple Difference in Differences

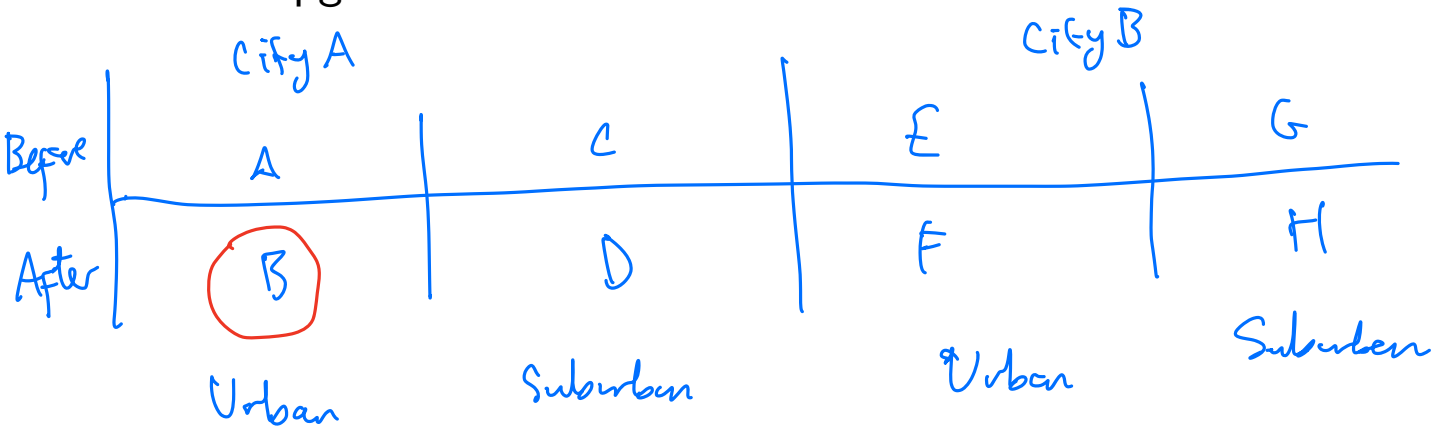
- Return to the store upgrades example. What if we discover all the upgraded stores were in urban areas, and non-upgraded stores were in the suburbs?
- We may no longer believe that changes in sales in the absence of upgrades would be the same. Recall:

$$\mathrm{E}\left(Y|G,T\right) = \beta_0 + \beta_1 G + \beta_2 T + \left[\beta_3 + ATT\right] G \cdot T,$$

  where the coefficient on $G \cdot T$ is the ATT iff common trends holds.
- What if we have data from stores in an entirely different city where no upgrades occurred?

Triple Difference in Differences

- Idea: Use difference in trends between urban and suburban stores in the city where no upgrades occurred to see what the difference in trends would have been in city A had there been no upgrades:



Triple Difference in Differences

- Identification of the ATT requires: "In the absence of treatment, the difference in the trend in sales across urban and suburban stores is the same in city B as it would have been in city A."
- Note: This allows for different trends between urban and suburban stores, but requires that difference is the same across cities.
- We now define the ATT as

$$ATT = \mathrm{E}\left(y\left(1\right) - y\left(0\right) | G = 1, T = 1, C = 1\right),$$

where $G = 1$ denotes urban stores, $C = 1$ denotes city A, and $T = 1$ denotes the time period after upgrades.

Implementation

$E[y(0)|G,T,C)$.

- Compute a diff-in-diff for city A, where treatment occured, and city B, where no treatment occurred, and difference them:

$$\begin{aligned}\mathrm{E}(Y|G,T,C) = \beta_0 + \beta_1 G + \beta_2 T + \beta_3 G \cdot T \\ + C(\gamma_0 + \gamma_1 G + \gamma_2 T + [\gamma_3 + ATT] G \cdot T).\end{aligned}$$

Under the assumption that the difference in trends is the same across the two cities, $\gamma_3 = 0$, so the triple diff-in-diff estimate of the ATT is the coefficient on $C \cdot G \cdot T$:

Implementation

- We get

$$\begin{aligned}\widehat{ATT} = &\left[\bar{Y}_{G=1,T=1,C=1} - \bar{Y}_{G=1,T=0,C=1}\right] \\ &- \left[\bar{Y}_{G=0,T=1,C=1} - \bar{Y}_{G=0,T=0,C=1}\right] \\ &- \left(\left[\bar{Y}_{G=1,T=1,C=0} - \bar{Y}_{G=1,T=0,C=0}\right]\right. \\ &\left.- \left[\bar{Y}_{G=0,T=1,C=0} - \bar{Y}_{G=0,T=0,C=0}\right]\right).\end{aligned}$$

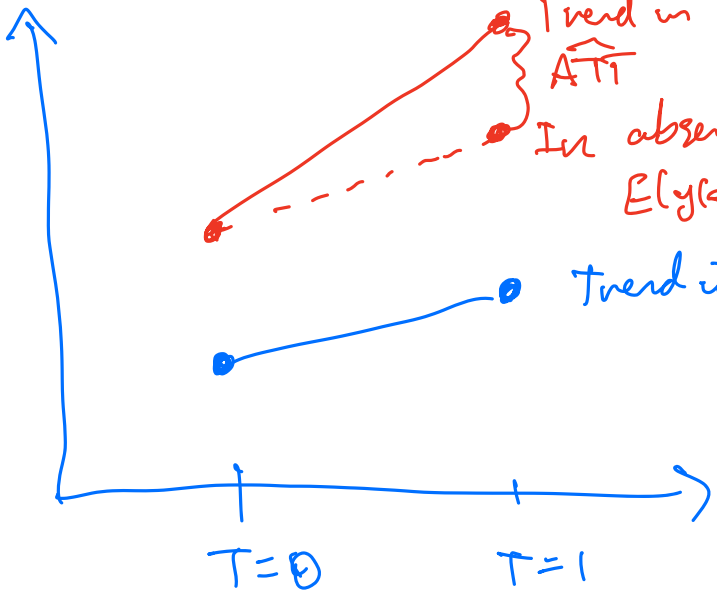
Diff in Diff for stores in A

$\longleftarrow$ DiD for stores in B.

Standard errors

- Computing standard errors assuming all observations are independent may lead us to have incorrect standard errors.
- Heteroskedasticity robust standard errors maintain the independence assumption.
- The concern is that the house price in 1978 will be correlated with its price in 1981, so treating these observations as independent is inappropriate.
- Often, cluster-robust standard errors are used, where a 'cluster' contains groups of observations that are likely to exhibit correlation over time, or within a particular subset of the cross section (say, at the state level).

Questions?



Trend in outcome for treated at $T=1$.

$\widehat{ATT}$

In absence of treatment, trend in $E[y(0)]$ is same for treated + untreated groups.

Trend in mean outcome untreated

This allowed us to obtain an unbiased est. of ATT.

Multiple Groups and Time Periods

- Suppose there are now multiple groups $g = 0, \ldots, G$, some of which are treated ($\mathcal{G}_1$) and some of which aren't ($\mathcal{G}_0$):

$$\mathcal{G}_0 = (0, \ldots, g_0 - 1); \quad \mathcal{G}_1 = (g_0, \cdots, G).$$

- We have group / individual level observations at time periods $t = 0, \ldots, T$:

$$\mathcal{T}_0 = (0, \ldots, t_0 - 1); \quad \mathcal{T}_1 = (t_0, \ldots, T).$$

Pre-treatment Post-treatment

- Treatment dummy given by

$$D = \mathbf{1}(t \geq t_0, g \geq g_0).$$

Saturated conditional mean

- Write a saturated specification for the conditional mean of $y_0|G, T$ as:

$$\mathrm{E}(y_0|G,T) = \alpha + \sum_{t=1}^{T} \beta_t \mathbf{1}(T=t) + \sum_{g=1}^{G} \gamma_g \mathbf{1}(G=g) + \sum_{g\geq 1, t\geq 1} \delta_{gt} \mathbf{1}(T=t, G=g).$$

Saturated spec. for conditional mean of y_0.

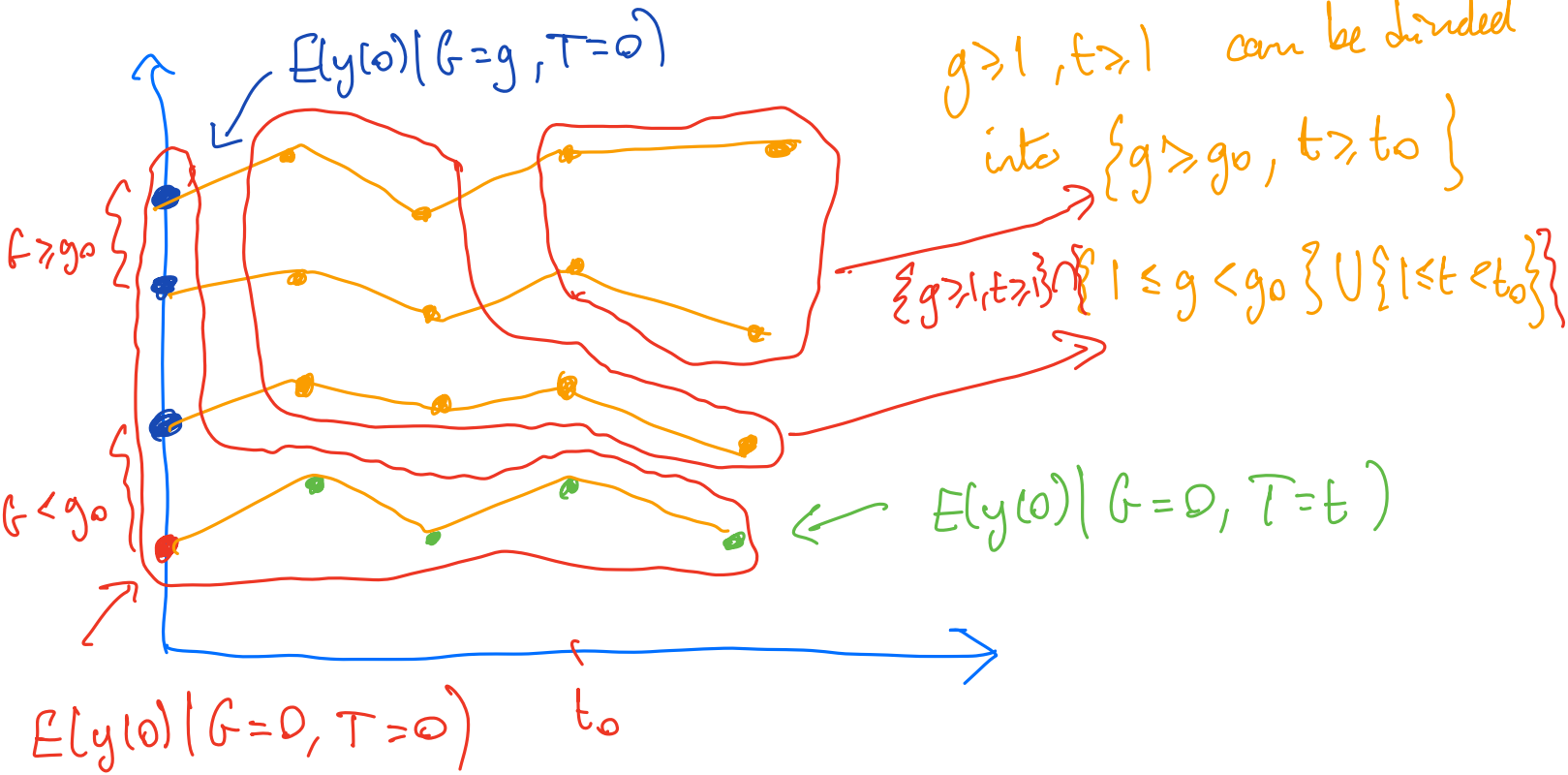
- Note that

$$\begin{aligned}\mathrm{E}(Y|G,T) &= \mathrm{E}(y_0|G,T) + D\mathrm{E}(y_1 - y_0|G,T) \\ &= \mathrm{E}(y_0|G,T) + \sum_{g\geq g_0, t\geq t_0} \pi_{gt} \mathbf{1}(T=t, G=g).\end{aligned}$$

$Y = y_0 + D(y_1 - y_0)$

All values of G, T where D=1.

Saturated conditional mean



Saturated conditional mean

- Writing all the terms out gives

$$\begin{aligned}\mathrm{E}(Y|G,T) = \alpha &+ \sum_{t=1}^{T} \beta_t \mathbf{1}(T=t) + \sum_{g=1}^{G} \gamma_g \mathbf{1}(G=g) \\ &+ \sum_{1\leq g<g_0 \cup 1\leq t<t_0,(g,t)\geq(1,1)} \delta_{gt} \mathbf{1}(T=t, G=g) \\ &+ \sum_{g\geq g_0, t\geq t_0} (\delta_{gt} + \pi_{gt}) \mathbf{1}(T=t, G=g).\end{aligned}$$

- Run regression on constant and fully interacted set of dummies.
- $\pi_{gt} = ATT_{g,t}$ identified if $\delta_{gt} = 0$. This means there is a common trend between $G = 0$ and $G = g$ between times $T = 0$ and $T = t$.

Common Trends

- Assuming common trends for all group in all time periods yields

$$\mathrm{E}(Y|G,T) = \alpha + \sum_{t=1}^{T} \beta_t \mathbf{1}(T=t) + \sum_{g=1}^{G} \gamma_g \mathbf{1}(G=g) + \sum_{g \geq g_0, t \geq t_0} \pi_{gt} \mathbf{1}(T=t, G=g).$$

- Further assuming treatment effect homogeneity across groups and time, so for all $t \geq t_0, g \geq g_0$:

$$\mathrm{E}(y_1 - y_0|G=g, T=t) = \mathrm{E}(y_1 - y_0|G \geq g_0, t \geq t_0)$$

yields

$$\mathrm{E}(Y|G,T) = \alpha + \sum_{t=1}^{T} \beta_t \mathbf{1}(T=t) + \sum_{g=1}^{G} \gamma_g \mathbf{1}(G=g) + \pi D.$$

Common Trends

"Two-way" fixed effects.

One-way in group

One-way in time.

$\pi \mathbb{1}(T \geq t_0, G \geq g_0)$

$= \pi D.$

Homogeneous TE.

- Having information from multiple control groups not helpful if we don't impose common trends in the pre-treatment period, since each group then has its own trend - which to compare with?
- Chaisemartin and Haultfœuille (2020) examine the estimand π in case treatment effect homogeneity is violated.
- Find that two-way fixed effects estimator estimates weighted averages of ATTs in treated groups, with potentially negative weights. Again, estimand may be negative even if all ATTs are positive.

Can estimate dynamic treatment effects with a homogeneity assumption on treatment effect per treated groups: $\pi_{gt} = \pi_t \ \forall g \geq g_0$.

Checking Pre-Trends

- Regression specification set $t = 0$ as the base year. $\delta_{gt} = 0$ for all t imposes common trends between $t = 0$ and $t \geq t_0$.
- If we add the assumption that common trends holds between *any* pre-period and post-period, we have a testable implication.
- Suppose $\mathcal{G} = \{0, 1\}$. Common trends implies for $1 \leq s < t_0$:

$$\begin{aligned} &\mathrm{E}\left(y_0 | G = 1, T = t_0\right) - \mathrm{E}\left(y_0 | G = 1, T = t_0 - s\right) \\ &= \mathrm{E}\left(y_0 | G = 0, T = t_0\right) - \mathrm{E}\left(y_0 | G = 0, T = t_0 - s\right). \end{aligned}$$

- For any choices $1 \leq r < s < t_0$, this implies

$$\begin{aligned} &\mathrm{E}\left(y_0 | G = 1, T = t_0 - r\right) - \mathrm{E}\left(y_0 | G = 1, T = t_0 - s\right) \\ &= \mathrm{E}\left(y_0 | G = 0, T = t_0 - r\right) - \mathrm{E}\left(y_0 | G = 0, T = t_0 - s\right). \end{aligned}$$

Checking Pre-Trends

Differential time trends between treated & control groups.

- These conditional means can be estimated, since they are untreated potential outcomes in the pre-periods.
- Check for $t < t_0$ whether $\delta_t = 0$ in

$$\mathrm{E}(Y|G,T) = \alpha + \gamma_1 + \sum_{t=1}^{t_0-1} \beta_t \mathbf{1}(T=t) + \sum_{t=1}^{t_0-1} \delta_t \mathbf{1}(T=t, G=1).$$

- Can use an F-test, but need to be careful of serial correlation and within-group correlation.

Conduct F test for $\delta_t = 0 \quad \forall t \geq 1$.

Standard errors

Individual i belongs to some group g. g is level at which treatment status is determined.

- If we have individual rather than aggregated data:

$$Y_{igt} = \alpha + \beta_t + \gamma_g + \pi D_{gt} + \epsilon_{igt},$$

  we can choose to cluster observations at different levels.
- At one extreme we assume data are iid across i and t, at the other we allow for correlation across all i with equal g and across all time periods.
- In the latter case, we effectively have as many observations as groups.

Clustered standard errors

- Suppose we have individual level data $\{Y_i, G_i, T_i, W_i\}_{i=1}^n$, and let X_i denote the included regressors (e.g. time/group interactions, covariates).
- We estimate

$$Y_i = X_i'\beta + U_i; \quad \mathrm{E}(X_i U_i) = 0.$$

using OLS and obtain

$$\hat{\beta} - \beta = \left(\sum_{i=1}^n X_i X_i' \right)^{-1} \left(\sum_{i=1}^n X_i U_i \right).$$

Clustered standard errors

- U_i can have arbitrary correlation within group and across time. Rewrite

$$\hat{\beta} - \beta = \left(\sum_{g=0}^{G} \left[\sum_{i:G_i=g} X_i X_i' \right] \right)^{-1} \left(\sum_{g=0}^{G} \left[\sum_{i:G_i=g} X_i U_i \right] \right)$$
$$= \left(\frac{1}{G+1} \sum_{g=0}^{G} \left[\sum_{i:G_i=g} X_i X_i' \right] \right)^{-1} \left(\frac{1}{G+1} \sum_{g=0}^{G} \left[\sum_{i:G_i=g} X_i U_i \right] \right).$$

- Let $n_g \leq M$ be the number of observations per cluster. If observations are independent across g:

$$Z_g = \sum_{i:G_i=g} X_i X_i'; \quad R_g = \sum_{i:G_i=g} X_i U_i$$

are independent across g.

Clustered standard errors

- Let

$$\Sigma_X = \lim_{G\to\infty} \frac{1}{G+1} \sum_{g=0}^{G} \mathrm{E}\left(Z_g\right)$$

$$\Omega = \lim_{G\to\infty} \frac{1}{G+1} \sum_{g=0}^{G} \mathrm{E}\left(R_g R_g'\right).$$

Need to be careful when defining limiting variance because clusters may have different numbers of observations.

- The distribution may differ across g in part because of different cluster sizes n_g.
- With $G \to \infty$, n_g fixed (and bounded):

$$\sqrt{G+1}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1}\right).$$

Clustered standard errors

- We haven't made any assumption about dependence within clusters.
- Eg. 1: iid data within and across clusters would mean:

$$\begin{aligned}\mathrm{E}\left(R_g R_g'\right) &= \sum_{i:G_i=g}\sum_{j:G_j=g}\mathrm{E}\left[U_i U_j X_i X_j'\right] \\ &= \sum_{i:G_i=g}\sum_{j:G_j=g;j\neq i}\mathrm{E}\left[U_i X_i\right]\mathrm{E}\left[U_j X_j'\right] \\ &+ \sum_{i:G_i=g}\mathrm{E}\left[U_i^2 X_i X_i'\right] = n_g\tilde{\Omega},\end{aligned}$$

where $\tilde{\Omega} = \mathrm{E}\left[U_i^2 X_i X_i'\right]$, so

$$\sum_{g=0}^{G}\mathrm{E}\left(R_g R_g'\right) = \tilde{\Omega}\sum_{g=0}^{G} n_g = O\left(G\right)$$

Clustered standard errors

- Eg. 2: All observations within cluster are equal, but iid across clusters:

$$\begin{aligned}
\mathrm{E}\left(R_g R_g'\right) &= \sum_{i:G_i=g} \sum_{j:G_j=g} \mathrm{E}\left[U_i U_j X_i X_j'\right] \\
&= \sum_{i:G_i=g} \sum_{j:G_j=g} \mathrm{E}\left[U_i^2 X_i X_i'\right] \\
&= n_g^2 \tilde{\Omega},
\end{aligned}$$

so

$$\sum_{g=0}^{G} \mathrm{E}\left(R_g R_g'\right) = \tilde{\Omega} \sum_{g=0}^{G} n_g^2 = O\left(G\right).$$

Large clusters

- If the number of observations per cluster (say n for each cluster for simplicity, $N = (G+1)\,n$) is growing, we get a different rate of convergence in the previous examples:
- iid within and across cluster:

$$\begin{aligned}\mathrm{E}\left(R_g R_g'\right) &= \sum_{i:G_i=g}\sum_{j:G_j=g}\mathrm{E}\left[U_i U_j X_i X_j'\right] \\ &= \sum_{i:G_i=g}\sum_{j:G_j=g;j\neq i}\mathrm{E}\left[U_i X_i\right]\mathrm{E}\left[U_j X_j'\right] \\ &\quad + \sum_{i:G_i=g}\mathrm{E}\left[U_i^2 X_i X_i'\right] = n\tilde{\Omega},\end{aligned}$$

- Perfectly correlated within cluster:

$$\begin{aligned}\mathrm{E}\left(R_g R_g'\right) &= \sum_{i:G_i=g}\sum_{j:G_j=g}\mathrm{E}\left[U_i U_j X_i X_j'\right] \\ &= n^2\tilde{\Omega}.\end{aligned}$$

Large clusters

- If the clusters are growing, we need a new rate of convergence.
- Under iid sampling

$$\sqrt{n(G+1)}\left(\hat{\beta}-\beta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1}\right),$$

where

$$\Sigma_X = \lim_{(n,G)\to\infty} \frac{1}{n(G+1)} \sum_{g=0}^{G} \mathrm{E}\left(Z_g\right),$$

$$\Omega = \lim_{(n,G)\to\infty} \frac{1}{n(G+1)} \sum_{g=0}^{G} \mathrm{E}\left(R_g R_g'\right).$$

Large clusters

- If observations are perfectly correlated within cluster:

$$\sqrt{(G+1)}\left(\hat{\beta}-\beta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1}\right),$$

where

$$\Sigma_X = \lim_{(n,G)\to\infty} \frac{1}{n(G+1)} \sum_{g=0}^{G} \mathrm{E}\left(Z_g\right),$$

$$\Omega = \lim_{(n,G)\to\infty} \frac{1}{n^2(G+1)} \sum_{g=0}^{G} \mathrm{E}\left(R_g R_g'\right).$$

Clustered standard errors

- Dependence within clusters can arise when the error U_i contains a cluster-specific component common to all units i ~~such that $i = g$~~: in cluster g:

$$U_{ig} = \lambda_g + \epsilon_{ij}.$$

- Clustering also arises in panel data because of serial correlation.
- Our application earlier likely features both - house prices are correlated over time and share common features with houses in the same place.

Cluster Covariance Estimator

- For n_g bounded, $G \to \infty$ we estimate the limiting variance as

$$\begin{aligned}\hat{V} &= \hat{\Sigma}_X^{-1}\hat{\Omega}\hat{\Sigma}_X^{-1} \\ &= \left(\frac{1}{G+1}\sum_{g=0}^{G} Z_g\right)^{-1}\left(\frac{1}{G+1}\sum_{g=0}^{G}\hat{R}_g\hat{R}_g'\right)\left(\frac{1}{G+1}\sum_{g=0}^{G} Z_g\right)^{-1},\end{aligned}$$

where $\hat{R}_g = \left[\sum_{i:G_i=g} X_i\hat{U}_i\right]$ and $\hat{U}_i = Y_i - X_i'\hat{\beta}$.

Cluster Covariance Estimator

- If $(n, G) \to \infty$ and there is perfect correlation in cluster, we would use the same $\hat{V}$:

$$\hat{V} = \left(\frac{1}{n(G+1)} \sum_{g=0}^{G} Z_g \right)^{-1} \left(\frac{1}{n^2(G+1)} \sum_{g=0}^{G} \hat{R}_g \hat{R}_g' \right) \left(\frac{1}{n(G+1)} \sum_{g=0}^{G} Z_g \right)^{-1},$$

while if the data are iid within cluster, replace with $n\hat{V}$:

$$n\hat{V} = \left(\frac{1}{n(G+1)} \sum_{g=0}^{G} Z_g \right)^{-1} \left(\frac{1}{n(G+1)} \sum_{g=0}^{G} \hat{R}_g \hat{R}_g' \right) \left(\frac{1}{n(G+1)} \sum_{g=0}^{G} Z_g \right)^{-1}.$$

Statistical Inference

- Which estimator of V to use? It converges at a different rate depending on the strength of intra-cluster dependence.
- Answer: It doesn't matter, since the resulting test statistic is the same:

$$\begin{aligned}
&(G+1)\cdot\left(\hat{\beta}-\beta\right)'\hat{V}^{-1}\left(\hat{\beta}-\beta\right)\\
&=\sqrt{G+1}\left(\hat{\beta}-\beta\right)'\hat{V}^{-1}\sqrt{G+1}\left(\hat{\beta}-\beta\right)\\
&=\sqrt{n(G+1)}\left(\hat{\beta}-\beta\right)'\left(n\hat{V}\right)^{-1}\sqrt{n(G+1)}\left(\hat{\beta}-\beta\right).
\end{aligned}$$

- Hence:

$$(G+1)\cdot\left(\hat{\beta}-\beta\right)'\hat{V}^{-1}\left(\hat{\beta}-\beta\right)\xrightarrow{d}\chi^2_p,$$

  where p is the dimension of X.
- Can use this limiting result to test linear restrictions on β in the usual manner.

Large number of observations in a cluster

- Different results are obtained if we assume there are n observations per cluster and $(n, G) \to \infty$ jointly.
- Provided the dependence within cluster is weak enough, the mean outcome aggregated at the group level is consistent for the group mean.
- Intuition: If all observations were iid, and the sample size is $n(G+1)$, we should expect:

$$\sqrt{n(G+1)}\left(\hat{\beta}-\beta\right) \xrightarrow{d} \mathcal{N}(0, V_A),$$

but if there is strong dependence within clusters, the rate of convergence ought to be $\sqrt{G+1}$:

$$\sqrt{(G+1)}\left(\hat{\beta}-\beta\right) \xrightarrow{d} \mathcal{N}(0, V_B).$$

- Can use same t/F stat in either case. What matters is $G \to \infty$. See C. Hansen (2007), Hansen and Lee (2019) for the general results.

Inference with few treated clusters

- In some applications, there is a single/few treated units. Our previous asymptotic approximation had the number of treated and control states going to infinity (clusters are randomly sampled).
- Conley and Taber (2011) assume the number of control units goes to infinity, deriving a limit distribution for the estimated treatment effect on the treated groups which is not consistent (since the number of treated groups stays fixed) but allows for inference with a limiting distribution for the estimator.
- Canay Santos Shaikh (2021) develop a bootstrap procedure when the number of clusters is small but the number of observations within a cluster is large. Their test controls size very accurately with inaccuracy decaying exponentially in number of clusters.