Implementing differences in differences

PSCI 2301: Quantitative Political Science II

Prof. Brenton Kenkel

brenton.kenkel@gmail.com

Vanderbilt University

April 7, 2025

Recap

Last time: Differences in differences to estimate causal effects

- Data requirements
  - → Observe same units over time
  - → Some units never treated, others sometimes treated
- Method to estimate the ATT
  1. Calculate diff over time among sometimes-treated obs
  2. Calculate diff over time among never-treated obs
  3. Subtract (2) from (1)
- Parallel trends assumption
  - → If treated units had not been treated, would have had same average trend over time as untreated ones
  - → Not directly testable (fundamental problem of causal inference)

Today's agenda

1. Using regression to estimate DiD
2. DiD in practice with the Getmansky, Grossman, Wright replication data
3. Wrap up with comparison of all the estimators we've studied

Regression for diff in diff

Things we already know about regression

If Z_i is binary and we run the bivariate regression

$$\mathbb{E}[Y_i \mid Z_i] = \alpha + \beta Z_i$$

...we get a difference in means.

Term	Notation	Interpretation
Intercept	α	average outcome when $Z_i = 0$
Slope	β	difference in averages b/w $Z_i = 1$ and $Z_i = 0$

Things we already know about regression

If Z_i is binary and we run the interacted regression

$$\mathbb{E}[Y_i \mid X_i, Z_i] = \alpha + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i)$$

... we get separate regression lines for the $Z_i = 0$ and $Z_i = 1$ groups.

Term	Notation	Interpretation
Intercept	α	intercept of X-Y relationship when $Z_i = 0$
X coefficient	β_1	slope of X-Y relationship when $Z_i = 0$
Z coefficient	β_2	shift in intercept when $Z_i = 1$
Interaction term	β_3	shift in slope when $Z_i = 1$

Regression for difference in differences

For each observation i in time period t, code the variables:

- $Y_{i,t}$: observed outcome for i at time t
- D_i: is i in the group that will get treated? (same across time periods)
- A_t: has the treatment been applied yet? (same across observations)

$$\mathbb{E}[Y_{i,t} \mid D_i, A_t] = \alpha + \beta_1 D_i + \beta_2 A_t + \beta_3 (D_i \times A_t)$$

- α: Pre-treatment avg in never-treated group
- β_1: Difference between groups in pre-treatment period
- β_2: Shift in post-treatment avg for never-treated group
- β_3: Difference in differences estimate

Why the interaction term gives the DiD estimate

$$
\begin{aligned}
DiD &= \text{diff over time in treated group} \\
&\quad - \text{diff over time in untreated group} \\
&= \{\mathbb{E}[Y_{i,t} \mid D_i = 1, A_t = 1] - \mathbb{E}[Y_{i,t} \mid D_i = 1, A_t = 0]\} \\
&\quad - \{\mathbb{E}[Y_{i,t} \mid D_i = 0, A_t = 1] - \mathbb{E}[Y_{i,t} \mid D_i = 0, A_t = 0]\} \\
&= [(\alpha + \beta_1 + \beta_2 + \beta_3) - (\alpha + \beta_1)] - [(\alpha + \beta_2) - \alpha] \\
&= (\beta_2 + \beta_3) - \beta_2 \\
&= \beta_3.
\end{aligned}
$$

Advantages of the regression approach

1. Can accommodate multiple time periods before/after
   - Just code $A_t = 0$ before treatment administered, $A_t = 1$ after
   - Include time trend (or dummies) to account for baseline changes:

$$\mathbb{E}[Y_{i,t} \mid \cdots] = \alpha + \beta_1 D_i + \beta_2 A_t + \beta_3 (D_i \times A_t) + \underbrace{\beta_4 t}_{\text{linear time trend}}$$

2. Can control for observable confounders
   - Include them in regression equation as normal:

$$\mathbb{E}[Y_{i,t} \mid \cdots] = \alpha + \beta_1 D_i + \beta_2 A_t + \beta_3 (D_i \times A_t) + \beta_4 X_{i,t}$$

   - Still need to avoid post-treatment bias, may need to use previous period measurements
3. Easier to calculate standard errors

Application to border walls and smuggling

Getmansky, Grossman, Wright data

```
df_ggw
```

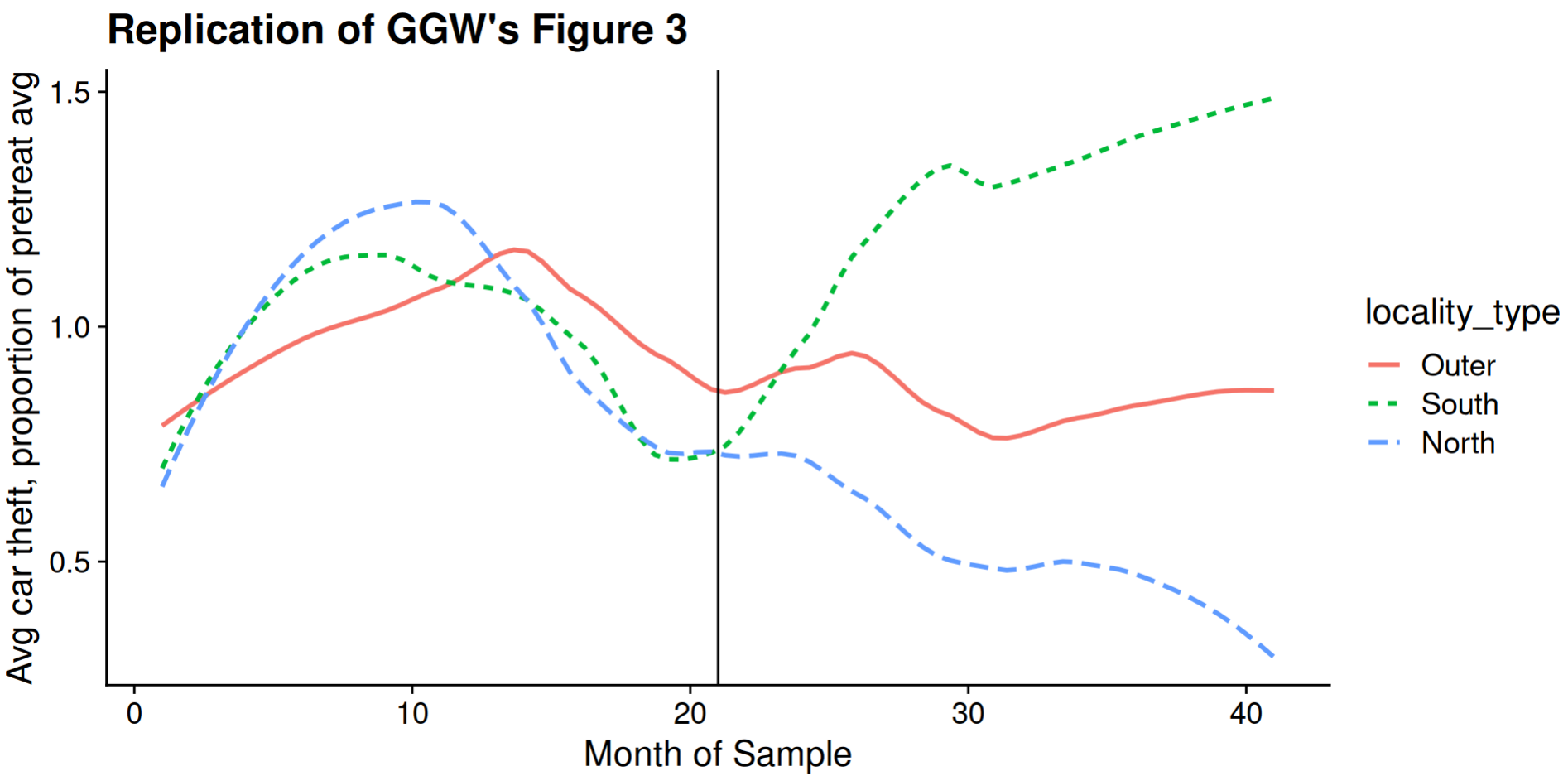
```
# A tibble: 37,010 × 206
  lcode locality_name PT        locality_type    TR   TRN   TRS Treatment MonthRun MonthFrom MonthTo
  <dbl> <chr>         <dbl+lbl> <fct>         <dbl> <dbl> <dbl>     <dbl>    <dbl>     <dbl>   <dbl>
1     7 SHAHAR        0 [Pre-b... South             0    NA     0         0        1         0      22
2     7 SHAHAR        0 [Pre-b... South             0    NA     0         0        2         0      21
3     7 SHAHAR        0 [Pre-b... South             0    NA     0         0        3         0      20
4     7 SHAHAR        0 [Pre-b... South             0    NA     0         0        4         0      19
5     7 SHAHAR        0 [Pre-b... South             0    NA     0         0        5         0      18
# i 37,005 more rows
# i 195 more variables: date <dbl>, month <dbl+lbl>, year <dbl>, district <dbl>, province <dbl>,
#   gaza <dbl>, wb <dbl>, urban <dbl>, distance2greenline <dbl>, religion <dbl+lbl>,
#   regional_council <dbl>, population <dbl>, carthefts <dbl>, breakins <dbl>,
#   carthefts_1molag <dbl>, breakins_1molag <dbl>, carthefts_1molead <dbl>, breakins_1molead <dbl>,
#   suicide_attacks <dbl>, suicide_killed <dbl>, suicide_injured <dbl>, killedi <dbl>,
#   injuredi <dbl>, attacks <dbl>, district_suicide_attacks <dbl>, district_suicide_killed <dbl>, ...
```

Assessing parallel trends

Code to replicate GGW's Figure 3

```
df_ggw |>
  # Calculate theft as proportion of pretreatment mean in each group
  group_by(locality_type) |>
  mutate(
    pretreat_mean = mean(carthefts_pc[MonthRun < 21], na.rm = TRUE),
    theft_prop = carthefts_pc / pretreat_mean
  ) |>
  # Calculate avg theft as prop of pretreat mean by group and month
  group_by(MonthRun, locality_type) |>
  summarize(avg_theft_prop = mean(theft_prop, na.rm = TRUE)) |>
  ggplot(aes(x = MonthRun, y = avg_theft_prop)) +
  geom_smooth(aes(linetype = locality_type, color = locality_type), span = 0.5, se = FALSE) +
  geom_vline(xintercept = 21) +
  labs(
    title = "Replication of GGW's Figure 3",
    x = "Month of Sample",
    y = "Avg car theft, proportion of pretreat avg"
```

Assessing parallel trends



Diff in diff “by hand”

```
df_ggw |>
  group_by(locality_type, PT) |>
  summarize(theft = mean(carthefts_pc, na.rm = TRUE)) |>
  summarize(theft_diff = mean(theft[PT == 1] - theft[PT == 0])) |>
  mutate(
    vs_south = theft_diff - theft_diff[locality_type == "South"],
    vs_outer = theft_diff - theft_diff[locality_type == "Outer"]
  )
```

```
# A tibble: 3 × 4
  locality_type theft_diff vs_south vs_outer
  <fct>              <dbl>    <dbl>    <dbl>
1 Outer            -0.0481   -0.253    0
2 South             0.205     0        0.253
3 North            -0.465    -0.670   -0.417
```

Diff in diff via regression

North vs South, no time trend or controls

- Repeated obs of same locality ⇝ cluster standard errors
- `estimatr::lm_robust()` does this better than what I'd used before

```
library("estimatr")
fit_simple <- lm_robust(
  carthefts_pc ~ locality_type * PT,
  data = df_ggw,
  subset = locality_type != "Outer",
  clusters = lcode
)

tidy(fit_simple) |>
  as_tibble() |>
  select(term, estimate, std.error)
```

```
# A tibble: 4 × 3
  term                   estimate std.error
  <chr>                     <dbl>     <dbl>
1 (Intercept)               0.866    0.0542
2 locality_typeNorth        0.158    0.113
3 PT                        0.205    0.0470
4 locality_typeNorth:PT    -0.670    0.0791
```

Diff in diff via regression

Adding a linear time trend

```
fit_time_lin <- lm_robust(
  carthefts_pc ~ locality_type * PT + MonthRun,
  data = df_ggw,
  subset = locality_type != "Outer",
  clusters = lcode
)

tidy(fit_time_lin) |>
  as_tibble() |>
  select(term, estimate, std.error)
```

```
# A tibble: 5 × 3
  term                    estimate std.error
  <chr>                      <dbl>     <dbl>
1 (Intercept)               0.856     0.0599
2 locality_typeNorth        0.158     0.113
3 PT                        0.186     0.0557
4 MonthRun                  0.000917  0.00230
5 locality_typeNorth:PT    -0.670     0.0791
```

Diff in diff via regression

Adding time dummies

```
fit_time_dummy <- lm_robust(
  carthefts_pc ~ locality_type * PT + factor(MonthR
  data = df_ggw,
  subset = locality_type != "Outer",
  clusters = lcode
)

tidy(fit_time_dummy) |>
  as_tibble() |>
  select(term, estimate, std.error) |>
  filter(!str_detect(term, "MonthRun"))
```

```
# A tibble: 4 × 3
  term                   estimate std.error
  <chr>                     <dbl>     <dbl>
1 (Intercept)             4.77e-1  6.09e- 2
2 locality_typeNorth      1.58e-1  7.33e- 3
3 PT                     -8.54e+9  1.33e+10
4 locality_typeNorth:PT  -6.70e-1  7.33e- 3
```

Diff in diff via regression

Adding controls: Population, urbanization, distance from Green Line (+ its square), whether local govt is a regional council

```
fit_time_full <- lm_robust(
  carthefts_pc ~ locality_type * PT + factor(MonthR
    popN + urban + distance2greenline +
    I(distance2greenline^2) + regional_council,
  data = df_ggw,
  subset = locality_type != "Outer",
  clusters = lcode
)

tidy(fit_time_full) |>
  as_tibble() |>
  select(term, estimate, std.error) |>
  filter(!str_detect(term, "MonthRun"))
```

```
# A tibble: 9 × 3
  term                      estimate std.error
  <chr>                        <dbl>     <dbl>
1 (Intercept)                 0.182    0.219
2 locality_typeNorth          0.0974   0.114
3 PT                          0.629    0.0941
4 popN                       -0.0140   0.0195
5 urban                      -0.181    0.188
6 distance2greenline          0.0466   0.00900
7 I(distance2greenline^2)    -0.00151  0.000219
8 regional_council            0.155    0.186
9 locality_typeNorth:PT      -0.671    0.0791
```

Diff in diff via regression

Full model for North vs Outer

```
fit_vs_outer <- lm_robust(
  carthefts_pc ~ locality_type * PT + factor(MonthR
    popN + urban + distance2greenline +
    I(distance2greenline^2) + regional_council,
  data = df_ggw,
  subset = locality_type != "South",
  clusters = lcode
)

tidy(fit_vs_outer) |>
  as_tibble() |>
  select(term, estimate, std.error) |>
  filter(!str_detect(term, "MonthRun"))
```

```
# A tibble: 9 × 3
  term                      estimate std.error
  <chr>                        <dbl>     <dbl>
1 (Intercept)               7.03e-1 0.211
2 locality_typeNorth        2.67e-1 0.143
3 PT                        3.17e-1 0.0901
4 popN                     -1.50e-2 0.00773
5 urban                    -1.32e-1 0.178
6 distance2greenline       -1.84e-2 0.00333
7 I(distance2greenline^2)   8.41e-5 0.0000164
8 regional_council          2.55e-2 0.158
9 locality_typeNorth:PT    -4.15e-1 0.0714
```

Interpretation

Table 1: Barrier construction and auto theft: deterrence and displacement.

	Diff-in-diff		
	North vs. South	North vs. Outer	South vs. Outer
Treatment	0.097	0.267*	0.211**
	(0.114)	(0.142)	(0.098)
Post	0.186***	0.102**	−0.230***
	(0.056)	(0.044)	(0.048)
Treatment × Post	−0.671***	−0.415***	0.256***
	(0.079)	(0.071)	(0.057)
N	24,985	23,716	25,069
Clusters	617	587	620

- Barrier construction reduced car theft on average
- ...but also displaced a lot from north to south

Wrapping up

Treatment effect estimation: Comparing the options

Method	Key assumptions	Pros	Cons
Difference in means	Random assignment	No fancy adjustments needed, small standard errors	Expensive or infeasible for many causal questions
Matching	All confounders measured	Easy to calculate and to assess balance	Many ways to match, curse of dimensionality, unlikely to measure all confounders
Regression	All confounders measured, linear relationship b/w them and outcome	Flexible, easy to interpret	Linear extrapolation can be problematic, unlikely to measure all confounders

Treatment effect estimation: Comparing the options

Method	Key assumptions	Pros	Cons
Instrumental variables	Instrument not confounded, doesn't directly affect outcome	Can leverage random influence on nonrandom treatment	Assumptions very stringent, standard errors large if instrument weak, effective sample (compliers) not necessarily representative
Regression discontinuity	Treatment "jumps" with running variable, confounders don't	Easy to assess balance, easy to see where causal estimate comes from	Possible sensitivity of estimate to bandwidth, effective sample (running $\approx$ 0) not necessarily representative
Difference in differences	Parallel trends in potential outcome if untreated	Widely applicable with panel data, easy to see where causal estimate comes from	Requires repeated observation of same units, parallel trends sketchy if other things changing when treatment is switched on

Plan for the rest of the semester

- Weds 4/9: A crash course on synthetic control
- Mon 4/14: Presentations of final projects
  - → Should be 10-12 minutes each
  - → Main points to hit: your causal question, your data, how you identify a causal effect, your main findings
- Weds 4/16: Likely no class
- Weds 4/23: Final paper and revision memo due