MY457: Problem Set 3 - Difference in Differences

Pedro Torres-Lopez, Michael Ganslmeier, Daniel de Kadt

Fri/01/Mar

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 12pm (noon) on Sat/09/Mar. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a study of the effect of a treatment $D_i \in 0, 1$ on Y_i for all $i \in 1, ..., N$. In this case, treatment is occurs across two dimensions: *i)* treatment group $G_i \in 0, 1$, and *ii)* time $t \in 0, 1$.

1.1. In this setting we can denote the following potential outcomes:

- $Y_{it}(0)$: potential outcome for unit i in period t when untreated
- $Y_{it}(1)$: potential outcome for unit i in period t when treated

Write out the realisations of these potential outcomes as observed data. Which are observed, when, and for which groups?

1.2 What is the main assumption in a canonical two-period difference-in-differences setting? Explain how violations of this assumption can impact the validity of the estimated treatment effect.

1.3 Given repeated cross-sectional data, we can estimate a canonical two-period difference-in-differences design with the following regresison specification:

$$Y_i = \hat{\alpha} + \hat{\gamma} D_i + \hat{\delta} T_i + \hat{\tau}(G_i * T_i) + \hat{\varepsilon_i}$$

Explain the parameter (estimand) that each coefficient in the specification estimates.

2 Simulations

In this question we will use simulated data to test some of our intuitions about difference-in-differences. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Explain the code below and relate it to a difference-in-differences data generating process. What kind of data (panel or repeated-cross sectional) is this?

```
set.seed(123)

n_units <- 1000

tau  <-  25000

G = rbinom(n_units, 1, 0.5)
for (i in 1:2) {
  data <- tibble(
    ID = 1:n_units,
    G  = G,
    T = ifelse(i == 2, 1, 0)
  )

  if (i == 1) {
    sim_data <- data
  } else {
    sim_data <- rbind(sim_data, data)
  }
}

Y0 <- rnorm(n_units, 50000, 2500)

data <- sim_data %>% mutate(
  Y0 = c(Y0, Y0*(1+1/10)),
  Y0 = ifelse(G == 1, Y0 + 10000, Y0),
  Y1 = Y0 + tau,
  Y = ifelse(G == 1 & T == 1, Y1, Y0)
)
```

2.2 Without using a regression, estimate the canonical two-period difference-in-differences using only `Y`, `G`, and `t`. What do you find?

2.3 Now estimate the difference-in-differences design using linear regression. Do you find any differences to your previous estimation? Why or why not?

2.4 Using the potential outcomes in our simulated data, create a plot visualizing the difference-in-differences estimator.

2.5 Now consider a new data generating process, given by the simulation code below. Explain how this code is different to the code in question 2.1.

```
set.seed(123)

n_obs <- 1000

n_periods <- 20

tau_values <- c(1000, 3000, 3000, 2000, 5000, 3000, 9000, 6000, 7000, 10000,
                9000, 8000, 6000, 3000, 7000, 2000, 5000, 2000, 1000)

tau <- setNames(tau_values, paste0("tau_", 1:19))

G = rbinom(n_obs, 1, 0.5)
```

```r
for (i in 1:20) {
  treated_units <- ifelse(i > 5, sample(1:n_obs, size = floor(1/40*n_obs)), NA)
  if (i == 1) {
    treated <- treated_units
  } else {
    treated <- c(treated, treated_units)
  }

  data <- tibble(
      ID = 1:n_obs,
      G  = G,
      P  = i,
      T  = ifelse(ID %in% treated, 1, 0)
    )

  if (i == 1) {
    sim_data <- data
  } else {
    sim_data <- rbind(sim_data, data)
  }
}

Y0 <- rnorm(n_obs, 50000, 2500)

sim_data <- sim_data %>%
  mutate(
    Y0 = (1 + P/10) * Y0 + if_else(G == 1, 10000, 0),
    Y1 = case_when(
      P %in% 1:19 ~ Y0 + tau[paste0("tau_", P)],
      TRUE ~ Y0
    ),
    Y = if_else(G == 1 & T == 1, Y1, Y0),
    D = T * G
  )

data <- sim_data
```

2.6 Using the new simulated data, estimate the difference-in-differences design using a two-way fixed effects linear regression. You can do this in multiple ways: using `lm` and `factor()`, using `lm` on de-meaned data, using `plm` with `model = "within"` and `effect = "twoways"`, or using `fixest`.

2.7 Using the new data and either the *fect* package or the *did* package, estimate dynamic period-specific ATTs and provide an event study plot. What do you find?

3 Replication

In this section, we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment.*

In recent decades, there has been an increasing push towards higher transparency in income, wealth, and earnings. Transparency facilitates comparisons between individuals. In 2001, Norwegian tax records became accessible online allowing individuals to have access to these easily, assuming they had access to internet.

The author uses this setting to analyze the effect of salary transparency on the subjective well-being of

individuals across the income distribution.

3.1 Read into `R` the replication data set (`Norway-MSD.dta`) and visualise the trend in Norwegian happiness (`po_happy`) over the years. Include a vertical line to indicate when treatment came into effect.

3.2 Explain, simply and in your own words, the causal inference problem faced by the authors (i.e., what confounding are they concerned about?). Then explain, simply and in your own words, the author's research design and how it mitigates the problems identified.

3.3 In what way is the author's design a difference-in-differences, and how does it differ from the cases we have typically seen in the lecture? Do you have any potential concerns about the plausibility of the underlying assumptions? You might benefit from reading section II of the paper closely.

3.4 Estimate the baseline specification as given in equation (1) in the paper. In addition to the difference-in-differences components, the regression should include a dummy variable for each year, and should control for marital status, education, household size, household workers, female, age and age squared. Hint: remember to include categorical variables as `factors()` where appropriate.

3.5 Estimate the same specification, but separately on two different subgroups in the data. First estimate the effect for those who have high access to internet, then for those who do not. Do you find any differences? What do you conclude from this exercise?

3.6 Test for parallel pre-trends using the event study design. What do you find?

3.7 (Extra credit): What do you think of the research design used in this paper? Do you have any suggestions for how it could have been improved, or extra falsification tests the author could have tried?