UNIVERSITY OF BAYREUTH

## Causal Inference

## Assignment 2

Due noon on Friday of Week 8

Word limit: 4000 words

## Exercise 1: Simple diff-in-diff

In this exercise, you will carry out the diff-in-diff analysis implied by Table XII of John Snow's *On the communication of cholera* (1855), which is reproduced below.[1] Snow was trying to establish that cholera was spread through contaminated water. In other parts of the study he presents evidence that nearby houses connected to different water supplies had different cholera risks, and that cholera outbreaks were concentrated around certain pumps providing public water. In this part he carries out an informal diff-in-diff, using the fact that the water source for some London sub-districts changed in 1852 when the Lambeth Company moved its intake from Hungerford Bridge (near Westminster) to Thames Ditton, 22 miles upstream. Snow argues that the cleaner water obtained in sub-districts served by the Lambeth Company after the change reduced the risk of death from cholera in these sub-district.

1. Create a panel dataset based on the table. Each row should be a subdistrict-year.

2. Report the mean number of deaths in each group in the pre and post periods; use these means to calculate the effect of the Lambeth Company's move assuming a valid diff-in-diff design.

3. State in words the *parallel trends assumption* that must be employed here. Is it likely to hold? Under what circumstances might it be problematically violated? What data would be useful to assess this assumption?

---

[1]  The book is available on Google Books.

4. Estimate the diff-in-diff (point estimate and standard error) using the **group diff-in-diff formulation, i.e. interaction formulation**.

5. Estimate the diff-in-diff (point estimate and standard error) using the **first differences formulation**.

6. Finally, estimate the diff-in-diff (point estimate and standard error) using the **fixed effects formulation**.

## TABLE XII.

| Sub-Districts. | Deaths from Cholera in 1849. | Deaths from Cholera in 1854. | Water Supply. |
|---|---|---|---|
| St. Saviour, Southwark . | 283 | 371 | |
| St. Olave   .   . | 157 | 161 | |
| St. John, Horsleydown . | 192 | 148 | |
| St. James, Bermondsey . | 249 | 362 | |
| St. Mary Magdalen   . | 259 | 244 | |
| Leather Market   .   . | 226 | 237 | Southwark & Vaux- |
| Rotherhithe*   .   . | 352 | 282 | hall Company only. |
| Wandsworth .   .   . | 97 | 59 | |
| Battersea   .   .   . | 111 | 171 | |
| Putney .   .   .   . | 8 | 9 | |
| Camberwell   .   .   . | 235 | 240 | |
| Peckham   .   .   . | 92 | 174 | |
| | | | |
| Christchurch, Southwark | 256 | 113 | |
| Kent Road   .   .   . | 267 | 174 | |
| Borough Road   .   . | 312 | 270 | |
| London Road   .   . | 257 | 93 | |
| Trinity, Newington   . | 318 | 210 | |
| St. Peter, Walworth   . | 446 | 388 | Lambeth Company, |
| St. Mary, Newington   . | 143 | 92 | and Southwark and |
| Waterloo Road (1st)   . | 193 | 58 | Vauxhall Compy. |
| Waterloo Road (2nd)   . | 243 | 117 | |
| Lambeth Church (1st) . | 215 | 49 | |
| Lambeth Church (2nd) . | 544 | 193 | |
| Kennington (1st)  .   . | 187 | 303 | |
| Kennington (2nd)   . | 153 | 142 | |
| Brixton   .   .   . | 81 | 48 | |
| Clapham   .   .   . | 114 | 165 | |
| St. George, Camberwell | 176 | 132 | |
| | | | |
| Norwood   .   .   . | 2 | 10 | |
| Streatham   .   .   . | 154 | 15 | Lambeth  Company |
| Dulwich   .   .   . | 1 | — | only. |
| Sydenham   .   .   . | 5 | 12 | |
| First 12 sub-districts   . | 2261 | 2458 | Southwk. & Vauxhall. |
| Next 16 sub-districts   . | 3905 | 2547 | Both Companies. |
| Last 4 sub-districts   . | 162 | 37 | Lambeth Company. |

# Exercise 2

A key problem incumbents encounter in civil wars is lack of information to combat insurgency. Given that insurgents exploit information asymmetries at the local level, they can easily hide and become a difficult target for incumbents. In the absence of such information, incumbents often resort to indiscriminate violence, via large-scale reprisals against entire villages suspected to host insurgents. One such example of indiscriminate violence is Aerial bombardment. Due to the nature of insurgency, bombing frequently occurs in and around settled areas, and consequently it tends to generate many civilian casualties. Using data from the Vietnam War, Kocher, Pepinsky and Kalyvas examine the effect of the bombings on Viet Cong support. In particular, they look at the impact of September 1969 bombings on hamlet control in December 1969. The dataset, `Vietnam_matching.dta`, is available on the course website.

## Data

The data comes from various sources. The United States compiled a gazetteer of South Vietnamese hamlets, identified their geographic coordinates, and conducted a census. District Senior Advisors (DSAs), Army officers ranking major or above, were assigned to complete detailed questionnaires, some on a monthly basis, others quarterly, for every village and hamlet in their zones of operation. DSAs, together with small American staffs, were detached from U.S. units to live and work in the districts they rated. The RVN had 261 districts with a median area of 377 kilometres squared, or about one-fourth the size of the median U.S. county. There was a median of 36 hamlets per district in 1969. One might analogize the problem a DSA faced to that of the sheriff of a small U.S. county trying to identify dangerous towns or neighborhoods in his or her jurisdiction. Linking bombings to hamlets, the authors construct a dataset that allows them within some margin of error to identify the number of bombings per hamlet in September 1969 and examine their impact on insurgency control. The variables of interest are as follows:

- `mod2a_1adec`: the "Enemy Military Model (2A)" (Hamlet Control), which rates the presence and activity of Viet Cong military units in the vicinity of each hamlet on a 5-point scale:"fully government controlled" (1), "moderately government controlled" (2), "contested" (3), "moderately insurgent controlled" (4), and "fully insurgent controlled.

- `bombed_969`: Number of bombings per hamlet in September 1969.

- `std`: Rough terrain

- `lnhpop`: log of hamlet population

- `ln_dist`: log distance from closest international boundary

- `score`: Development index score

- `mod2a_1ajul`: Enemy Military Model (2A) in July 1969

- `mod2a_1admn`: District average control before September 1969

Your task is as follows:

1. Use a matching estimator to derive the effect of September bombing in September 1969 on insurgency control in December 1969. Choose matching estimator other than Caliper matching, but describe how you have chosen the estimator and upon which assumptions it rests. To conduct the analysis, operationalise the treatment as a binary variable. Discuss which covariates you use in the matching procedure and why.

2. Assess balance in pre-treatment covariates between treated and control units, before and after matching.

3. Use Capiler matching using a capiler of 0.25 and estimate the treatment effect. Assess balance before and after Caliper balancing between the control group and the treated group. Is Caliper matching doing a better job than the algorithm you chose in the previous question? Which matching procedure do you prefer and why?

4. Estimate the effect of the number of bombings in September 1969 on hamlet control in December 1969 using a multivariate regression by conditioning on covariates. Discuss which covariates from the dataset you have included and why. Does the point estimate and standard error you estimate change from the ones you have found in 1.3.? Discuss which result you find more convincing and why.

Analyses that rely on random number generators should include the following seed: set.seed(02022022).

# Exercise 3: Value of Connections in Politics

Should policy making and government officials listen to the opinion of NGOs and industry members and should they be under the influence of external bodies and experts? On the one hand, we can imagine that listening to the private sector and other non-governmental bodies might allow them to have more accurate information and hence make better policy choices. Yet, this could also have negative consequences, especially if we imagine that the influence external bodies have on government officials can be bought. Lobbyists (paid advocates who aim to influence the decisions of legislators or other government officials) play an increasingly important role in the political system of the United States and other democracies. Does lobbying entail a fair playing ground?

Lobbying firms attract individuals leaving government positions with high-paying offers and in exchange gain access to important politicians still in office. By doing so, these lobbying firms hope to obtain favourable legislation and to receive inside information on what is happening in government. This has led to concerns that corporations and other organisations are able to buy influence and acquire privileged access to important politicians, and by so doing, undermine democratic values.

In this exercise, we want to understand to what extent former government officials can monetise on the personal connections acquired during their period of public service in Congress.

**Question a)** Imagine you are interested in estimating the effect of personal political connections on the amount of revenue received from lobbying contracts.

$$\text{Revenue}_{i,t} = \alpha + \beta \text{Connections}_{i,t} + u_{i,t}$$

The main dependent variable of interest is (log) revenue per lobbyist. The main independent variable of interest is *connections* which is calculated as the count of currently serving politicians a lobbyist is linked to through their previous employment experience.

- Under what assumptions could we interpret $\beta$ as a causal estimate?

- What could be the source of selection bias?

- Discuss how the various methods learnt in the past weeks (multivariate analysis, matching, panel

fixed effects and difference-in-differences) could be applied to this question. What would be your preferred method?

**Question b)** Imagine we run a regression of revenue on time dummies and a dummy for "ever connected to Senator" and "ever connected to Representative".

- What is the reason to include time dummies?

- Why do the Senate/Representative dummies help isolate our effect of interest?

- Would it be reasonable to put in individual politician served (senator, representative) fixed effects in this regression? If yes, what variation would that control for?

**Question c)** A crucial policy question is understanding whether the effect of connections is driven by the effects of expertise (or ability) of the lobbyist on earnings, or these result from acquiring political connections while serving a politician. We might imagine that the first mechanism doesn't undermine democratic values, yet the second one does and could invite introducing restrictions on employment after holding public office. How can we use the existing data to test whether the effect is driven by lobbyists' ability or from acquiring connections?

- How would you address this problem? - indicate a model specification.

- Given your specified model, what assumptions do we need to make to estimate the causal effect of connections on revenue?