PROBLEM SET 2

Causal Inference (HT 2025)

Instructions:

You must submit two files –

1. A PDF file with your write-up and results, including neatly formatted tables and figures.
2. An R Mardkdown file that generates this PDF, and contains your code for analysis.

Ensure that files are named xx_PS2.pdf and xx_PS2.Rmd, where xx is your candidate number. Also specify your candidate number in the main body of both files. Do not include other identifying information, such as your name, college, etc.

Collaborating with other students is permitted, under three conditions:

1. You must first try to answer each problem on your own before meeting with classmates.
2. Every student is required to produce their own code and write-up. No copy-pasting!
3. You must indicate who you collaborated with: students working together should come up with a random five digit number and mention this group ID in their PDF submissions.

Problem 1: Institutions and Economic Development

Total points: 50

The assignment is based on the famous Acemoglu, Johnson & Robinson (AJR) 2001 study on the importance of inclusive institutions for economic development. [1] AJR argue that institutions leave a long imprint on countries' economic activity. They distinguish between inclusive and extractive institutions. The former diffuses economic returns across different strata of society, whereas the latter facilitates the appropriation of wealth by elites.

To provide evidence for the importance of institutions, they turn to the colonial structures of the 19th and early 20th centuries. Their identification strategy comes from variation in geography and climate, which determined whether colonizers would establish inclusive or extractive institutions. In those areas in which settlers encountered high mortality rates, they built extractive institutions without long-term planning. In areas with low mortality rates, they built inclusive institutions. Assuming that settler mortality rates satisfy the assumptions of an instrumental variable, this allows the authors to identify the causal effect of institutions on economic development over the long term. The data can be found on Canvas as AJR.dta. The key variables are the following:

- **logpgp95**: the logged GDP per capita measured in 1995. This is Y , the dependent variable of interest. (Logging is a typical way of bringing a variable's distribution closer to normal. For background information on the log transformation, have a look at https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf).
- **avexpr**: Average protection against expropriation risk (1985—1995). This is an index of how extractive a given set of institutions is. This is our treatment variable D, which is, of course, not randomly assigned.
- **logem4**: Logged settler mortality rates. This is Z, our instrument.

You can also use other variables from the dataset if you think this helps in any part of your analysis. Your task is to answer the following questions:

[1] Acemoglu, Johnson, & Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." American Economic Review, 91(5): 1369-1401.

a) Is there a link between institutions and economic development? This is not a causal question; we are asking if there is any association between the two. Provide a scatterplot to show this is the case.

b) Is this relationship causal? How do mortality rates help in answering this question?

c) Estimate the ITT and interpret it.

d) Estimate the LATE, using both a Wald estimator and a 2SLS estimator. Interpret your findings.

e) Assess the plausibility of the IV assumptions in this setting. For each of the assumptions listed below, explain (in words) what it means substantively in the context of this study and provide a statistical test or verbal argument assessing its plausibility.

- Relevant first stage
- Monotonicity
- Independence
- Exclusion restriction

Problem 2: Centralization and Public Investment

Total points: 50

This question uses data from Malesky, Nguyen, and Tran's (2014) study of the effect of centralizing control of public services in Vietnam. [2] Scholars have long speculated that de-centralized control of public services in developing countries can lead to corruption and 'capture' by local politicians who fail to use tax revenue to invest in public services. Centralization of control may, therefore, lead to increased public investment. This is difficult to test because control of government spending rarely changes hands. But in 2009, Vietnam decided to experiment with re-centralizing control of its public services in some areas but not others, placing them under central government control instead of control by local district authorities in around one-fifth of districts nationwide. The authors use a difference-indifferences framework to ask what impact this had on infrastructure spending, comparing changes in treated districts (where spending was re-centralized) to untreated districts (where it remained under local control). They have data on infrastructure spending before the change (in 2006 and 2008) and after the change (in 2010), for small areas called 'communes', which are subsets of districts.

The dataset for this exercise is malesky.Rda and contains the following variables by commune:

- **year**: 2006,2008 or 2010
- **treatment**: =1 if the commune is in a treated district, 0 otherwise
- **infra**: index measuring infrastructure investment ranging from 0 to 5: higher values indicate greater investment
- **lnpopden**: population of the commune, logged
- **district**: district ID number

Your task is to answer the following questions:

a) Create a new dataset for the years 2008 and 2010 only. In this dataset, create: (i) a dummy variable called "post" equalling 1 if the year is after the treatment and (ii) a variable for the interaction between post and treatment.

[2] Edward Malesky, Cuong Viet Nguyen and Anh Tran (2014). "The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam." American Political Science Review 108 (1), pp. 144-168.

b) Use the dataset and variables you created in a) to calculate a difference-in-difference estimate of the causal effect of centralization on infrastructure investment. Interpret the resulting coefficient and its statistical significance.

c) Difference-in-differences estimation relies on the 'parallel trends' assumption. Explain what that assumption means in this study.

d) Now, we'll assess the parallel trends assumption empirically. Return to the full dataset and calculate the means of the outcome variable for 2006, 2008, and 2010 for both the treated and control groups (six means in total). Plot these means separately over time for the treated and control groups. Do you think that the parallel trends assumption holds in this case?

e) Create a new dataset for the years 2006 and 2008 only and use it to estimate a placebo difference-in-differences effect before the treatment occurred. What do you conclude about the parallel trends assumption?