POLI 171: Problem Set 5: Regression and Matching

Due Tuesday, March 13, at the beginning of class

For this assignment, you may collaborate with one other student. You may also complete this assignment on your own, if you prefer. If you collaborate, you and your co-author will turn in a single document, and you will each receive the same grade. You and your co-author are expected to work together on *all* parts of this assignment; do not simply divide the questions between yourselves.

Please type your answers and submit them as a hardcopy. Please show your work by including your `R` code and output in your answers.

Recentralization and Infrastructure Provision (7 points)

Like most countries today, Vietnam has multiple levels of government, and the responsibilities of the local governments steadily increased during the late 20th century. At the district level, the most important authority is the District People's Council (DPC), which has the power to elect or dismiss the district executive, the members of the district courts, and a variety of other district-level offices. The DPCs also make policy at the local level; they maintain their own budgets, and they are responsible for approving socioeconomic development plans each year. The DPCs are directly elected by citizens, but the Communist Party of Vietnam vets all candidates. In practice, participation in district elections is low, and the DPCs in most districts are "captured" by local elites whose interests often differ from those of the central government.

In 2007, the Communist Party of Vietnam passed a resolution calling for the abolition of the DPCs on the grounds that they were too inefficient and corrupt. However, the central government was nervous about the potential adverse side-effects of abolishing the councils, and it decided to begin the "recentralization" process in only a few test districts. In 2009, the Ministry of Home Affairs chose 10 provinces (containing 99 districts) where the recentralization process would begin later that year, and the districts in all other provinces remained decentralized. The recentralized districts were intended to be representative of the broader set of districts, and they included districts from all regions of the country and a variety of socioeconomic conditions. Nevertheless, treatment assignment was **not** random, and the selection of recentralization districts was likely driven by political considerations. The DPCs in the recentralized districts were abolished in April, 2009, and their powers were transfered up one level, to the Provincial People's Committees. The DPCs in all other districts continued to function as they had before.

In this part of the problem set, you will use a differences-in-differences model to estimate the effect of recentralization (the abolition of a DPC) on infrastructure provision, a policy area that is a high priority for the central government but tended to be a low priority for the district government. This problem is based on the article "The Impact of Recentralization on Public Services: A

Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam," by Edmund Malesky, Cuong Viet Nguyen and Anh Tran.

Question 1

Table 1 shows the proportion of communes that had paved, year-round roads over time, by the treatment status of their district. Communes are the lowest levels of government, and they are subordinate to the district government.

Table 1: Proportion of Communes with a Good Road in 2006, 2008 and 2010, by Treatment Status

Commune Type	2006	2008	2010
Remained Decentralized after 2009	0.79	0.80	0.80
Recentralized in 2009	0.77	0.75	0.84

The top row shows the proportion over time for the communes whose district remained decentralized after 2009 (the control group), and the bottom row shows the proportion over time for the communes whose district's DCP was abolished in 2009 (the treatment group).

a) One of the most common ways that researchers assess whether the parallel trends assumption holds is by examining the treatment and control groups' time trends prior to treatment. Based on Table 1, do you think that the treatment and control communes were experiencing a similar time trend prior to treatment? You are free to argue either position, but your answer should reference the specific numbers in this table and demonstrate that you understand what the parallel trends assumption is.

b) Use the data in Table 1 to calculate the differences-in-differences estimage of the treatment effect of recentralization on the proportion of communes that have good roads. For the pre-treatment period, use only the data from 2008 (not the data from 2006). Make sure that you show your work. Interpret the treatment effect in words.

Question 2

The `Vietnam` dataset contains commune-level data on treatment status and infrastructure outcomes. The following table describes the variables that are in this dataset.

Variable	Description
`Province`	Province ID
`RecentralizedCommune`	= 1 if the commune belongs to a district whose local council was abolished in 2009, and = 0 otherwise
`After`	= 1 if the observation comes from 2010, and = 0 if the observation comes from 2008
`GoodRoad`	= 1 if the commune has a paved road that is usable throughout the year, and = 0 otherwise

The observations are commune-years. Note that we do not have commune identifiers in the dataset, so we are unable to track the same commune across time (in fact, we do not even have true panel data in this case). Nevertheless, it is still possible to use this sort of data to calculate a differences-in-differences estimate because we have the treatment status of each commune and we also have Province IDs (the level of government at which treatment was assigned).

a) Use this dataset and the `lm()` function in R to calculate the differences-in-differences estimate of the effect of recentralization on the likelihood that a commune has a good road. Include Province fixed effects in your model, and cluster the standard errors at the province level. Interpret the treatment effect in words. How does it compare to the treatment effect that you calculated in Question 1(b)?

Head Start Technical Assistance and Child Mortality (7 points)

Head Start is a collection of early childhood programs that was created in 1965, as part of LBJ's Great Society. The eligibility requirements vary from program to program, but Head Start was designed to target young children in low-income households. The original components of Head Start (prior to its expansion in the 1980s) were early childhood education (originally summer school, but later year-round school), social services, health services, and nutrition. Although the funding is provided by the federal government, the initial roll-out of Head Start programs in the late 1960s was very decentralized: local organizations such as schools and public health agencies had to organize their own local program and apply for funding grants from the federal government.

Some policy-makers were concerned that local organizations in the poorest parts of the country would lack the experience or administrative capabilities necessary for writing successful grant applications, and that they would thus lose out in the nation-wide grant competition. In order to prevent this from happening, the Presidential Management Intern Program sent interns to the 300 poorest counties in the United States (counties that had a poverty rate above 59.1984% in the 1960 Census). These interns lived in their assigned county for around a month, and during that time they identified local actors who might be able to manage a Head Start program in the county, they encouraged these actors submit grants for Head Start programs, and they helped them write the grants. This grant-writing assistance was given only to the 300 poorest counties; other counties did not receive federal grant-writing assistance, though they were still allowed to apply for grants.

In this part of the problem set, you will use a regression discontinuity design to calculate the treatment effect of the Head Start grant-writing assistance on 5 to 9 child mortality in the 1970s. This problem is based on the article "Does Head Start Improve Children's Life Chances?" by Jens Ludwig and Douglas Miller. For this problem, you will use the `HeadStart` dataset, which includes the following variables:

Variable	Description
`CountyID`	County ID code
`State`	State ID code
`PovertyRate.1960`	County's poverty rate in the 1960 census
`PercentBlack.1960`	The percent of the county's population that identified as African American in the 1960 census
`PercentUrban.1960`	The percent of the county's population that lived in an urban area, as measured in the 1960 census
`LogPopulation.1960`	The log of the county's population, as measured in the 1960 census
`ChildMortalityRate.1960s`	The county's average mortality rate of children between the ages of 5 and 9, measured during the early 1960s. The units are deaths per year per 100,000 children in that age rage.
`ChildMortalityRate.1970s`	The county's average mortality rate of children between the ages of 5 and 9, measured during the 1970s.

Question 3:

a) Identify the running variable and the treatment threshold in this case.

b) Test for balance on pre-treatment variables by running placebo Regression Discontinuity tests on the pre-treatment variables `PercentUrban.1960`, `PercentBlack.1960`, `LogPopulation.1960` and `ChildMortalityRate.1960s`. Perform these placebo tests using the `rdrobust()` function, and report the results in a neatly-formated table that includes the treatment effect (the estimated difference in means at the cutoff) and the corresponding robust p-value.

c) Use the `rdrobust()` function to calculate a regression discontinuity estimate of the treatment effect on child mortality in the 1970s. Interpret the treatment effect in words. Is it statistically significant?

d) Display the treatment effect graphically by making a RD plot using the `rdplot()` function. In words, interpret what this graph is showing. What is the relationship between the running variable and child mortality overall? What happens at the treatment threshold?

e) Based on the description of treatment assignment in this question prompt, how likely is it that counties sorted around the treatment threshold or selected into treatment by controlling their value of the running variable? For this subquestion, I am not asking you to perform a statistical test; you will do that in the next subquestion. Instead, I want you to base your answer on an argument about the practicality of sorting into treatment in this case. What would have to happen in order for sorting to occur? Who do you think would make the decision to sort into or out of treatment? What information would the relevant decision-makers have to know in order to ensure that their county would get treated? What characteristics of their county would they have to control? Is it likely or unlikely that they would be able to do this in this particular case?

f) Test for clumping or sorting around the treatment threshold by performing a McCrary density test. Report the p-value and interpret the results of the test in words. What is the null hypothesis? Do you reject the null hypothesis? Based on this test, is there evidence of sorting or clumping around the threshold?