003: IV + RD, Robust

EC 607

Due *before* midnight on Sunday, 30 May 2021

DUE Your solutions to this problem set are due *before* 11:59pm on Sunday, 30 May 2021 on Canvas.

Your problem set **must be typed** with R code beneath your responses. *E.g.*, `knitr` and `R Markdown`.

OBJECTIVE This problem set has three purposes: (1) reinforce the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics.

An application to WTP for environmental quality

README These data come from the 2008 paper *Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program* by Michael Greenstone and Justin Gallagher. The paper attempts to estimate how consumers value local environmental quality (and its associated benefits/costs). To answer this question, Greenstone and Gallagher ("G&G") apply a *hedonic pricing model*, which bascially says the value of a property (house) is equal to the sum of the values of the propery's pieces—including the house's attributes, its neighborhood, its school, the property taxes, local environmental quality, *etc.* In other words: Holding all else constant, the hedonic model for a house's value could tell us by how much the house's value would change if the air got a bit cleaner (or if we added a fifth bathroom).

Your analysis will use four Stata files: `allsites.dta`, `allcovariates.dta`, `sitecovariates.dta`, and `2miledata.dta`. You can read Stata files into R using the `read_dta()` function from `haven`. I did not create these files—you get to see what "real" research files look like.

The data are at the level of US Census tract. Each of the US's ~65,000 Census tracts has about 4,000 people. The housing data are from the dicennial censuses, 1970–2000.

We're going to focus on how a tract's change in housing values between 1980 and 2000 was affected by clean ups of hazardous waste sites (plus 1980 demographic variables). Variation in clean ups of hazardous waste sites comes from the US EPA's Superfund Program. The EPA uses the *Hazardous Ranking System* (HRS) to determine which of the waste sites are the "worst". Only the worst sites get tested. Finally, sites whose HRS score exceeds 28.5 are placed on the *National Priorities List* (NPL) and are eventually cleaned up.

1.1 Load the `allsites.dta` data using `read_dta()` from `haven`. Now view it using the `View()` function. You should see nice variable labels. Get to know the data a bit. Describe anything that sticks out to you. `glimpse()` is a handy function for getting to know a dataset.

1.2 Let's start simple. Regress the 2000 log median house value for a Census tract on the indicator for whether the tract was listed on the National Priorities List before 2000. Report your results, interpreting the coefficient of interest.

1.3 Does clustering your errors by state affect your standard errors? Do you think it should? Explain.

Hint: The `felm()` function from `lfe` allows you to cluster and is quite fast.

1.4 Now run the three regressions described below, sequentially adding more controls to the regressions you ran in **1.2** and **1.3**. We are still principally interested in the effect of listing on the NPL.

- Control for 1980 housing values
- Also control for economic and demographic variables. (Report which variables you included.)
- Also add state fixed effects.

Briefly interpret your results.

1.5 Under what conditionsl will the coefficients on the NPL indicator (in **1.2**–**1.4**) be unbiased for the effect of NPL listing on local property values?

1.6 Let's compare the covariates for the treated (NPL listed) and control (non-NPL listed) tracts. First use `allcovariates.dta` to compare the covariates (the controls you used above) between the tracts listed on the NPL and the tracts that were not listed on the NPL (before 2000).

Does it look like the covariates are *balanced* across NPL and non-NPL listed tracts?

Notes: The `all` in the filename means that the file has *all* of the Census tracts. This comparison should be done via regression. You do not need to cluster your errors (though in real life you would want to).

1.7 Repeat the exercise from **1.6** with the dataset named `sitecovariates.dta`. This dataset focuses on tracts that received an HRS test in 1982. Separately compare the balance for treated and control units using the following three definitions of treated and control:

- NPL-listed prior to 2000 *vs.* not NPL-listed prior to 2000 (what we've been doing so far)
- HRS score above 28.5 *vs.* HRS score below 28.5
- HRS score in [16.5, 28.5) *vs.* HRS score in [28.5, 40.5]

What are your conclusions from these three comparisons?

1.8 Suppose we want to instrument NPL listing with HRS score. What assumptions are necessary?

1.9 Imagine we instead want to estimate the effect of NPL listing using a regression discontinuity where HRS score is the running variable and the cutoff is 28.5. What assumptions are necessary?

1.10 Consider the following three "facts":

- The EPA states that the 28.5 cutoff was selected because it produced a manageable number of sites.
- None of the individuals involved in identifying the site, testing the level of pollution, or running the 1982 HRS test knew the cutoff/threshold.
- EPA documentation emphasizes that the HRS test is an imperfect scoring measure.

Do these three facts suggest that the requirements for IV and/or RD as valid or invalid? Explain.

1.11 Now using the `2miledata.dta` dataset: Estimate the effect of NPL listing on log median house value (in 2000) using HRS score as an instrument. Estimate the three standard regressions of 2SLS (first stage, reduced form, second state) and briefly discuss them. Cluster your errors at the state level.

Make sure that your second-stage estimate is equal to the ratio of your reduced form and the first stage.

1.12 Does adding state fixed effects to the 2SLS estimates in **1.11** change your estimates? Use the first-stage and reduced-form estimates to explain what is happening. Do you think adding additional controls could be important?

1.13 Repeat the 2SLS analysis from **1.11** but change your instrument to an indicator for whether HRS score is above 28.5 (no fixed effects). How do you results change? Briefly discuss.

1.14 Based upon your first stage estimates in **1.13**: If we want to estimate the effect using a regression discontinuity, will it be a *sharp or fuzzy RD*? Briefly explain your answer.

1.15 Create the standard plots of a regression discontinuity (remember to bin your observations using the running variable, HRS score):

- The outcome variable *vs.* the running variable
- The treatment variable *vs.* the running variable
- Covariates *vs.* the running variable
- Bin counts *vs.* the running variable

1.16 Based upon your figures in **1.15**, does this look like a plausible regression discontinuity? Use the figures to explain your answer.

1.17 Time for RD estimates! Limit your analysis to HRS scores between [16.5, 40.5] and re-estimate the 2SLS from **1.13**. Add controls if you think they're necessary. Report your results.

1.18 Which of your estimates do you find most credible? Explain.