PPHA 42200: Problem Set 2

Professor: Koichiro Ito

March 27, 2022

Note: Datasets are changed from previous years.

Part I: Theory Questions (RD, Matching)

1. Suppose that the density of a running variable is smooth at a RD threshold. This is evidence that manipulation of the running variable did not occur. Answer TRUE/FALSE/Uncertain and briefly explain why.

2. Is the continuity of conditional regression functions—$E(Y_{0i}|X_i = x)$ and $E(Y_{1i}|X_i = x)$ are continuous in x— a) a necessary condition, b) a sufficient condition, or c) a necessary & sufficient condition for estimating a treatment effect in a sharp RDD? Choose either a, b, or c and briefly explain why.

3. For the following questions, consider an OLS regression equation, $Y = \alpha+\tau D+\beta X+\gamma X\cdot D+u$, for a sharp RD design, with a binary treatment D, an outcome variable Y, a running variable X, and an additive unobservable term u, where $D = 1$ if $X \geq 0$ and 0 otherwise.

   (a) Suppose that true values of β and γ are $\beta > 0$ and $\gamma > 0$ and that you forgot to include $X_i \cdot D_i$ as a control variable. Is it likely to create upward/downward/no bias for the estimate of τ, or is this uncertain? Explain why

   (b) Suppose that true values of β and γ are $\beta = 0$ and $\gamma = 0$. In this case, the RD estimate recovers the ATE for the entire sample. Answer true/false/uncertain and explain why.

   (c) Suppose that the conditional expectation for the untreated potential outcome, $E[Y_0|X]$, is discontinuous at some value of X. This violates the identification assumption of the sharp RD design and biases the estimate. Answer true/false/uncertain and explain why.

4. There is a binary treatment D, an outcome variable Y, and an observable variable $X = \{0, 1, 2\}$. The treatment effect is homogeneous conditional on X. Suppose that (Y_0, Y_1) is independent of D conditional on X, and $E[D|X]$ is linear in X. In this case, a cell estimator based on the three cells $X = \{0, 1, 2\}$ is a better method than OLS regression that includes X as the only control variable. Answer true/false/uncertain and explain why.

5. Suppose that $p(X_i)$ is the propensity score of a treatment D_i. Assume the two assumption required for the propensity score matching discussed in class are satisfied.

(a) Prove that:

$$\begin{aligned} ATE &\equiv E[Y_{1i}] - E[Y_{0i}] \\ &= E\left[\frac{D_i Y_i}{p(X_i)}\right] - E\left[\frac{(1-D_i)Y_i}{1-p(X_i)}\right] \\ &= E\left[\frac{(D_i - p(X_i))y_i}{p(X_i)(1-p(X_i))}\right] \end{aligned}$$

(b) Similarly, show that:

$$\begin{aligned} ATET &\equiv E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E\left[\frac{(D_i - p(X_i))y_i}{Pr[D_i = 1](1-p(X_i))}\right] \end{aligned}$$

(c) Describe the Conditional Independence Assumption (CIA) and explain a circumstance in which you would use matching and defend the assumption. Give the defense.

For questions (a) and (b), try to prove them by yourself first. If you get stuck, see page 876 of Cameron and Travedi or section 18.3.2 of Wooldridge to get help.

Part II: Empirical Analysis (RD)

A few notes on the the Empirical Analysis questions below: You will be using a modified version of the data sets, so your results will not exactly match those in the paper. In order to make sure your estimates match the solutions, you can use the following variables

- housing characteristics for 1(a):

  firestoveheat80 noaircond80 nofullkitchen80 zerofullbath80 bedrms* blt*

  detach80occ mobile80occ

- economic and demographic variables for 1(a):

  pop_den8 shrblk8 shrhsp8 child8 old8 shrfor8 ffh8 smhse8 hsdrop8 no_hs_dipl*

  ba_or_b* unemprt8 povrat8 welfare8 avhhin8 tothsun8 ownocc8 occupied80

- control variables for 3a (and again in 4):

  firestoveheat80_nbr noaircond80_nbr nofullkitchen80_nbr zerofullbath80_nbr

  bedrms* blt* detach80occ_nbr mobile80occ_nbr pop_den8_nbr shrblk8_nbr

  shrhsp8_nbr child8_nbr shrfor8_nbr ffh8_nbr smhse8_nbr hsdrop8_nbr no_hs_dipl*

  ba_or_b* unemprt8_nbr povrat8_nbr welfare8_nbr avhhin8_nbr tothsun8_nbr

  ownocc8_nbr occupied80_nbr

The empirical portion of this problem set is based on the paper: Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program (Greenstone and Gallagher 2008). The goal of the paper is to test the Market Willingness to Pay WTP for hazardous waste site cleanup using housing prices. Specifically, the paper is interested in observing how housing prices respond to changes in local environmental quality and the associated health risks. The analysis uses a hedonic property value model which assumes that a single housing price can be thought of as the sum of prices for all of the housing and neighborhood characteristics That is, holding all housing and neighborhood characteristics constant, if one were to add a 2-car garage onto a house, then the market price of the house will increase by an amount equal to how much the marginal consumer values the added garage. In the case of an environmental "bad", such as living in close proximity to a hazardous waste site, one would expect that the market price of nearby housing would increase if the hazardous waste site were cleaned up (all else equal). This problem set will give you a chance to test this hypothesis.

You will use 4 different data files in your analysis. Three of the data files (allsites.dta, allcovariates.dta, and sitecovariates.dta) will be used in part (1) of this question for background analysis. The 4th file (2miledata.dta) will be used for most of the empirical analysis. The variables in the files should be clear from either their names or their labels.
The unit of observation in all data files is a US Census Tract (roughly 4,000 people and 1,000 housing units). There are roughly 65,000 US Census Tracts. Those census tracts with missing property value housing data were dropped. All of the census tract housing data are from the 1970-2000 Decennial Censuses. We are most interested in the change in housing values between 1980 and 2000. To simplify matters for the purposes of this problem set we only include 1980 housing characteristics and property values from 1980 and 2000. The paper compares the evolution of housing prices between census tracts where hazardous waste sites were and were not cleaned up. The paper also looks at

distances around the hazardous waste site (e.g. 1-4 miles) and uses the average housing values in concentric circles (of 1-4 miles) as the dependent variable. In this paper most of the analysis will focus on a data file using the "2 mile" sample.

The variation in environmental quality comes from US EPA's Superfund Program. The goal of the Superfund program is to cleanup the country's worst hazardous waste sites. There is a process for identifying the worst sites which includes a site assessment and 2 levels of site inspection (where soil and water samples are taken). If a site is among the most polluted or dangerous then scientists apply the Hazardous Ranking System (HRS) test on the site. Those sites that receive an HRS test are thought to be the worst of the worst (less than 1% receive a test). The test is a way to get a composite score for the site. Those sites that pass a certain score threshold (28.5 out of 100) are placed on the "National Priorities List" (NPL). Once on the NPL the site is legally obligated to be cleaned up. We will abstract from hazardous waste site characteristics for the purpose of this problem set.

The important thing to note is that the HRS score provides a potential opportunity to apply a Regression Discontinuity research design. The Superfund program began in 1980. The first sites were tested using the HRS in 1982 and those sites that scored above 28.5 were the first sites listed on the NPL in 1983. One variable we focus on is whether a census tract contains a hazardous waste site listed on the NPL by year 2000. You can think of the list of NPL 2000 sites as representing some weighted combination of those sites that are slated for cleanup (where site cleanups are in different stages of completion including some that were not started as of 2000).

1. This question asks you to run OLS regressions that look at whether there is an association between 2000 housing values and whether a census tract contained a hazardous waste site that was placed on the NPL by 2000.

   (a) Use the file allsites.dta. This file contains only own tract housing variables (i.e. no 2 mile averages). Use "robust" standard errors for all regressions. First regress 2000 housing prices on whether the census tract had an NPL site in 2000. Include 1980 housing values as a control. Next add housing characteristics as controls. Run a third regression adding economic and demographic variables as controls. Finally run a 4th regression that also includes state fixed effects. Briefly interpret the regressions. Under what conditions will the coefficients on NPL 2000 status be unbiased?

   (b) Here we will compare covariates between potential treatment and comparison groups. First, use allcovariates.dta to compare covariates (i.e. those used in the above regressions) between census tracts with and without a hazardous waste site listed on the NPL by 2000. Next, use sitecovariates.dta to compare covariates between those census tracts with a hazardous waste site that had an HRS test in 1982. Specifically, compare those with sites that scored above 28.5 to those that scored below 28.5. Finally, compare those census tracts with sites between 16.5 and 28.5 to census tracts with sites between 28.5 and 40.5. What conclusions do you draw from these 3 comparisons?

2. This question examines the possibility of using a Regression Discontinuity research design. Note that the rest of the empirical question will use the file 2miledata.dta. The housing variables in this file are 2 mile averages.

   (a) Consider the HRS score as the running variable for an RD research design. What assumptions are needed on the HRS score? How do each of the following "facts" impact the

appropriateness of these assumptions: (i.) The EPA assertion that "the 28.5 cutoff was selected because it produced a manageable number of sites." (ii.) None of the individuals involved in identifying the site, testing the level of pollution, or running the 1982 HRS test knew the cutoff threshold score.(iii.) EPA documentation emphasizes that the HRS test is an imperfect scoring measure.

(b) Create a histogram of the distribution (i.e. density) of the 1982 HRS scores by dividing the HRS score into non-overlapping bins. Include a vertical line at 28.5. Next, run local linear regressions on either side of 28.5 using the midpoints of the bins as the data. What do you conclude?

3. This question examines the 1st stage equation of an RD design using the 1982 HRS score.

   (a) Use a 2SLS (IV) econometric setup that uses whether or not a census tract has a site scoring above/below 28.5 as the instrument. Write down the 1st stage equation. Run the 1st stage regression experimenting with the same set of covariates used in question (1). In addition, run a second specification in which you limit the sample to only those census tracts with sites between 16.5 and 40.5 and run the specification using all of the control variables (we will use this as the size of the bandwidth for the "regression discontinuity" regression). Interpret the results.

   (b) Create a graph plotting the the 1982 HRS score against whether a site is listed on the NPL by year 2000 (NPL on the y-axis, HRS on the x -axis). Briefly explain and interpret this graph.

   (c) Create a graph that plots the 1982 HRS score against 1980 property values (property values on the y-axis, HRS on the x -axis). What do you conclude from this graph?

4. Write down the 2nd stage equation (with housing values as the outcome) and the 2 standard assumptions for valid IV estimation. Run 2SLS to get the estimated coefficient on 2000 NPL status. Run the same two specifications as in the previous question. Briefly interpret the results.

5. Write a 1 paragraph conclusion summarizing your findings and interpreting the results. Be sure to comment on how the evidence from this problem set supports the primary research question.