# ARE 213 PS 3

S. Sung, H. Husain, T. Woolley, A. Watt

2021-12-8

## Contents

# Problem 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.**

## Part (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?**

Table 1: Preliminary Hedonic Regressions

| | Log(Median 2000 Housing Prices) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| NPL 2000 | 0.0334*** | 0.0396*** | 0.0679*** | 0.0625*** |
| | (0.0125) | (0.0119) | (0.0101) | (0.0092) |
| Mean 1980 Housing Values | Yes | Yes | Yes | Yes |
| Housing Characteristics | No | Yes | Yes | Yes |
| Economic Vars and Demographics | No | No | Yes | Yes |
| State Fixed Effects | No | No | No | Yes |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Robust Standard Errors in parentheses.
NPL 2000 = has site on the National Priorities List by year 2000.

If we focus on regression (4), the coefficient on NPL 2000 is telling us that the marginal house buyer purchasing a house of median value in the year 2000 is willing to pay 6 percentage points higher housing price if a super fund site in the census tract has been labeled for cleanup using the NPL list.

The NPL 2000 coefficient will be unbiased if the NPL superfund assignment is conditionally independent of determinants of housing values in 2000, conditional upon covariates included in the regression. For example, regression (1) will be unbiased if, given mean housing values from 1980, the NPL assignment is as-good-as-random. Regression (4) would be unbiased if, in a given state, the NLP assignment is as-good-as-random after conditioning on 1980 mean housing values and observed housing, economic, and demographic characteristics of the census tract.

## Part (b)

**Here we will compare covariates between potential treatment and comparison groups. First, use allcovariates.dta to compare co-variates (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?**

# Problem 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.**

## Part (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:**

**—— a.i**

**The EPA assertion that the 28.5 cutooff was selected because it produced a manageable number of sites."**

**—— a.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.**

**—— a.iii**

**EPA documentation emphasizes that the HRS test is an imperfect scoring measure.**

## Part (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?**

# Problem 3

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

## Part (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.**

## Part (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.**

## Part (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?**

## Problem 4

**Write down the 2nd stage equation (with housing values as the out-come) 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.**

## Problem 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.**

## Appendix A: R Code

```r
rm(list=ls())
knitr::opts_chunk$set(echo = F)
# stargazer table type (html, latex, or text)
# Change to latex when outputting to PDF, html when outputting to html
table_type = "latex"

# install.packages("Synth")
library(tidyverse)
library(haven)
library(stargazer)
library(ggplot2)
library(tinytex)
# library(Synth)
# library(plm)
library(lmtest)
library(sandwich)
# library(gridExtra)
# library(grid)
# library(gtable)
# library(fastDummies)
# library(EnvStats)
# Load all sites data
data1a = read_dta('allsites.dta') %>%
    # select columns to omit from regressions
    select(-bedrms0_80occ, -blt0_1yrs80occ, -detach80occ) %>%
    # Drop duplicated rows
    distinct()

# Create fips lookup table (need to fill in fips = 32, 28, 46)
state_lookup = data1a %>%
    group_by(statefips) %>%
    select(statefips, state) %>%
    slice(which.max(nchar(as.character(state))))

# Plot vars that might be collinear
# scaleFUN <- function(x) sprintf("%.10f", x)
# data1a %>%
#     mutate(tot_blt = blt0_1yrs80occ+blt2_5yrs80occ + blt6_10yrs80occ +
#                 blt10_20yrs80occ + blt20_30yrs80occ + blt30_40yrs80occ + blt40_yrs80occ,
#             tot_attach = detach80occ + attach80occ + mobile80occ) %>%
#     ggplot() + geom_histogram(aes(x=tot_attach)) + scale_x_continuous(labels=scaleFUN)
# Regress prices on NLP, 1980 prices
reg1a1 = lm(lnmdvalhs0 ~ npl2000 + lnmeanhs8, data=data1a) %>%
    coeftest(vcov = vcovHC(., type = "HC0"))

# Regress prices on NLP, 1980 prices, housing characteristics
reg1a2 = data1a %>%
    select(lnmdvalhs0, npl2000, lnmeanhs8, tothsun8:occupied80) %>%
    lm(lnmdvalhs0 ~ ., data=.) %>%
    coeftest(vcov = vcovHC(., type = "HC0"))
```

```r
# Regress prices on NLP, 1980 prices, housing, econ, demographics
reg1a3 = data1a %>%
    select(-fips, -state, -statefips) %>%
    lm(lnmdvalhs0 ~ ., data=.) %>%
    coeftest(vcov = vcovHC(., type = "HC0"))

# Regress prices on NLP, 1980 prices, housing, econ, demographics, state FE
reg1a4 = data1a %>%
    select(everything(), -fips, -state, statefips) %>%
    mutate(statefips = factor(statefips)) %>%
    lm(lnmdvalhs0 ~ ., data=.) %>%
    coeftest(vcov = vcovHC(., type = "HC0"))

stargazer(reg1a1, reg1a2, reg1a3, reg1a4,
          title = "Preliminary Hedonic Regressions",
          dep.var.caption = "Log(Median 2000 Housing Prices)",
          dep.var.labels.include = FALSE, model.names = FALSE,
          # column.labels = c("FE", "FE(Cluster)"),
          keep = "npl2000",
          covariate.labels = 'NPL 2000',
          add.lines=list(c('Mean 1980 Housing Values', 'Yes', 'Yes', 'Yes', 'Yes'),
                         c('Housing Characteristics', 'No', 'Yes', 'Yes', 'Yes'),
                         c('Economic Vars and Demographics', 'No', 'No', 'Yes', 'Yes'),
                         c('State Fixed Effects', 'No', 'No', 'No', 'Yes')),
          font.size = "footnotesize", column.sep.width = "1pt", no.space = TRUE, omit.stat=c("f", "ser"
          type = table_type, header = FALSE, digits = 4,
          notes = c('Robust Standard Errors in parentheses.',
                    'NPL 2000 = has site on the National Priorities List by year 2000.'))
```