ARE 213 PS 3

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

2021-12-8

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 coeffcients on NPL 2000 status be unbiased?

```
##
## ================================================================
##                                  Dependent variable:
##                      ------------------------------------------
##                                                lnmdvalhs0
##                            coefficient           felm
##                                test
##                        (1)      (2)      (3)           (4)
## ----------------------------------------------------------------
## npl2000              0.033*** 0.034*** 0.068***      0.063***
##                      (0.013)  (0.012)  (0.010)       (0.009)
##
## ----------------------------------------------------------------
## Observations                                          42,881
## R2                                                    0.779
## Adjusted R2                                           0.779
## Residual Std. Error                                0.294 (df = 42793)
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

First, note that before starting Q1, we observed duplicate observations in 'allsite.dta'. These duplicates were dropped.

Second, we can see that coefficient on NPL2000 was statistically significant under all four specifications.

In order to be consistent and unbiased, we need NPL2000 to be randomly assigned (or as good as random given controls for more complex model specifications) and that there exists no correlation between NPL2000 and 2000 Housing Price via unobservable factor, which would thereby be captured through an error term.

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?

```
##            variable p_val_sp1 t_val_sp1 p_val_sp2 t_val_sp2 p_val_sp3 t_val_sp3
## 1       attach80occ     0.000    -9.089     0.041    -2.056     0.298    -1.044
## 2           avhhin8     0.000    -5.395     0.013     2.493     0.486     0.698
## 3     ba_or_better8     0.000   -10.898     0.000     4.915     0.036     2.114
## 4     bedrms0_80occ     0.047    -1.991     0.458     0.742     0.641     0.467
## 5     bedrms1_80occ     0.063    -1.863     0.617    -0.500     0.897    -0.130
## 6     bedrms2_80occ     0.001     3.376     0.001    -3.354     0.041    -2.057
## 7     bedrms3_80occ     0.490     0.690     0.636     0.474     0.472     0.721
## 8     bedrms4_80occ     0.004    -2.858     0.000     4.106     0.089     1.707
## 9     bedrms5_80occ     0.000    -4.184     0.010     2.574     0.219     1.232
## 10   blt0_1yrs80occ     0.004    -2.873     0.042     2.035     0.960     0.051
## 11 blt10_20yrs80occ     0.033     2.136     0.025     2.248     0.687     0.404
## 12   blt2_5yrs80occ     0.517     0.649     0.116     1.574     0.994    -0.008
## 13 blt20_30yrs80occ     0.114     1.581     0.520     0.643     0.961     0.049
## 14 blt30_40yrs80occ     0.811     0.239     0.060    -1.884     0.481    -0.707
## 15   blt40_yrs80occ     0.000    -4.098     0.007    -2.705     0.896    -0.130
## 16  blt6_10yrs80occ     0.000     3.885     0.018     2.381     0.649     0.456
## 17           child8     0.000     7.771     0.958     0.053     0.568     0.571
## 18      detach80occ     0.556    -0.589     0.050     1.968     0.108     1.618
## 19              ffh8     0.000    -6.931     0.018    -2.389     0.862     0.174
## 20  firestoveheat80     0.000     4.267     0.814    -0.236     0.510    -0.660
## 21          hsdrop8     0.623     0.492     0.235    -1.189     0.303    -1.033
## 22      mobile80occ     0.000     8.832     0.793    -0.263     0.285    -1.072
## 23   no_hs_diploma8     0.000     5.697     0.000    -4.679     0.060    -1.890
## 24      noaircond80     0.000     7.598     0.253    -1.145     0.870    -0.164
## 25  nofullkitchen80     0.089     1.701     0.402    -0.839     0.787    -0.271
## 26       occupied80     0.000     4.570     0.940     0.076     0.989    -0.014
## 27          ownocc8     0.000     7.040     0.922    -0.098     0.244    -1.169
## 28         pop_den8     0.000   -44.804     0.068    -1.833     0.571    -0.568
## 29          povrat8     0.044    -2.018     0.109    -1.606     0.716     0.364
## 30          shrblk8     0.000    -3.666     0.037    -2.095     0.926     0.093
## 31          shrfor8     0.000    -6.628     0.735    -0.339     0.785    -0.273
## 32          shrhsp8     0.000    -5.288     0.841    -0.201     0.928    -0.090
## 33           smhse8     0.000     6.358     0.001    -3.283     0.244    -1.168
## 34         tothsun8     0.029     2.187     0.951    -0.062     0.576    -0.561
## 35         unemprt8     0.005     2.832     0.001    -3.386     0.731    -0.345
## 36         welfare8     0.239    -1.178     0.041    -2.050     0.578    -0.557
## 37   zerofullbath80     0.009     2.625     0.089    -1.704     0.386    -0.868
```

In the table (technically saved in dataframe format), we check for balance of covariates (same ones used in 1(a)) across three potential treatment and control groups. In the first specification, treatment group is tracts on NPL list by 2000 and control is the counterpart. In the corresponding column 1 and 2, we can see that means are statistically significantly different on most variables across the two groups. In the second specification, treatment group is census tracts with HRS score in 1982 above 28.5 and control goups is census

tracts with score below 28.5. And the tracts without HRS testing in 1982 are omitted. Comparing across 37 covariates, we can again see that on 16 covariates there is statistically significant difference between the two groups. In the final specification, we get closer to treatment and control groups we would use in Regression Discontinuity design. We can see that p_values are significantly larger and we seeem to observe balance across most covariates. Hence, except for which tracts were placed on NPL list or not, the two groups look similar on average on observable dimensions.

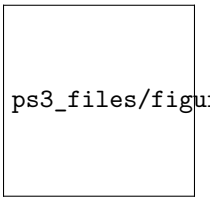


Table 1:

Statistic	N	Mean	Min	Max
npl2000	48,176	0.019	0	1

Statistic N Mean Min Max

treat 487 0.628 0 1

Statistic N Mean Min Max

treat 227 0.604 0 1

The three love plots above graphically depict the standardized difference of means and variance ratios between the NPL2000 groups. The rule of thumb is a standardized difference of means smaller than 0.25 in magnitude, ideally less than 0.1 and as close to 0 as possible. The rule of thumb for variance ratios is less than a factor of 2 (so between 0.5 and 2 in magnitude).

The first love plot is comparing the census tracts with and without a hazardous waste site listed on the NPL by 2000 (48,176 sites total, with 1.9% of them on the NPL). The second love plot comparing census tracts with a hazardous waste site that had an HRS test in 1982, above and below the 28.5 threshold (which is a significant reduction in sample size: 487 tracts with 62.8% above the threshold). We can see the difference in means and variance ratios for all our variables are getting closer to the rule of thumb above.

The last love plot has census tracts with a hazardous waste site that had an HRS test in 1982 with an HRS score between 16.5 and 40.5, comparing tracts above and below the 28.5 threshold. This further restricts the sample to 227 tracts with 60.4% above the threshold. The standardized difference of means are

not substantially different from the second plot but the variance ratios have gotten worse for some of our covariates.

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:

The assumptions we need to hold for a cutoff to be useful in an RD setting are that (1) The probability of treatment (NPL) should change discontinuously at the threshold of running variable (HRS). (2) We should not see bunching around the threshold, which may be indicative of manipulation in treatment status.

Assumption (1) is not an huge issue in this setting since the cutoff was strict in one direction. The fact that is is not strict in the other direction, however, makes us think it could be a slight concern. The following could either support assumption (2) or call it into question.

—— **a.i**

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

This might cause one to question assumption (2) because it introduces a sense that 28.5 was chosen intentionally rather than at random. Imagine, for instance, that if the regulators are selecting the cutoff based on manageability, then the regulators might line up all of the projects by their scores and determine the cutoff at a cleanup-cost discontinuity.

—— **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.

If true, then the data were not gathered with any knowledge of the cutoff ex ante, which make assumption (2) a little more likely to hold.

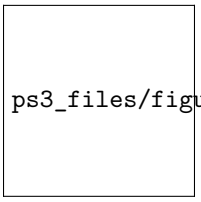
—— **a.iii**

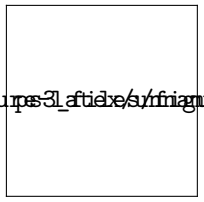
EPA documentation emphasizes that the HRS test is an imperfect scoring measure.

If the HRS test is an imperfect measure, then this also supports assumption (2). The more imperfect HRS is, the more comparable are the treated and untreated groups and therefore the more "randomly" the threshold is likely to have been assigned.

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?






```
## integer(0)
```

In the first histogram, there seems to be difference between regression intercepts. However, in the second histogram, when we restrict the local regressions to only include histogram bars closer to the threshold (including 16.5 to 40.5), the difference between the regression intercepts seems to disappear. The confidence intervals also show that there is no discontinuity in the density of HRS scores at the HRS score threshold, which suggests that there was no gaming of the system. This makes the RD design more credible in the restricted sample.

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.

Equations for 2SLS Set Up 1st Stage:

$$NPL2000_{i,s} = \alpha_1 + \beta_{1,1} 1(HRS1982 > 28.5)_i + \beta_{1,2} Covariates_i + \beta_{1,3}\theta_s + v_{i,s}$$

2nd Stage:

$$HousingPrice2000_{i,s} = \alpha_2 + \beta_{2,1} \hat{NPL}_{i,2000} + \beta_{2,2} Covariates_i + \beta_{2,3}\theta_s + \epsilon_{i,s}$$

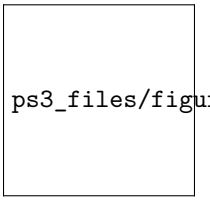
Note: θ_s is state fixed effects.

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                npl2000
##                          (1)             (2)
## ------------------------------------------------
## ind_hrsabove           0.799***        0.703***
##                        (0.034)         (0.054)
##
## ------------------------------------------------
## Observations             483             226
## R2                      0.797           0.705
## Adjusted R2             0.759           0.566
## Residual Std. Error 0.228 (df = 407) 0.301 (df = 153)
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

From this, we can conclude that NPL2000 is extremely predictive in the first stage! This implies that we were wise to use the IV approach before running the second stage reduced form.

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.

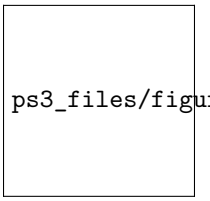


Above graph demonstrates that we have a clear break at threshold but the cut-off is not as perfect for strict RD design. Rather, we may want to consider Fuzzy RD design. The few census tracts that has lower than

28.5 points on HRS score in 1982 measure but still manages to get on NPL by 2000 could perhaps (1) receive higher score on on HRS in proceeding years, getting onto NPL before 2000, or (2) could potentially have other reasons, i.e. political connections or pressure, for getting onto the list.

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?



While median house price is definitely correlated with HRS score, it seems to be continuous around the threshold and therefore satisfies our assumptions for RD design.

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.

Equations for 2SLS Set Up 1st Stage:

$$NPL2000_{i,s} = \alpha_1 + \beta_{1,1}1(HRS1982 > 28.5)_i + \beta_{1,2}Covariates_i + \beta_{1,3}\theta_s + v_{i,s}$$

2nd Stage:

$$HousingPrice2000_{i,s} = \alpha_2 + \beta_{2,1}\hat{NPL}_{i,2000} + \beta_{2,2}Covariates_i + \beta_{2,3}\theta_s + \epsilon_{i,s}$$

Note: θ_s is state fixed effects.

The standard assumptions for valid IV estimation is (1) Cov(HSR1982, $\epsilon\_\{i, s\}) = 0 and (2) Cov(HSR1982, NPL2000) = / = 0$.

```
##
## ===================================================
##                            Dependent variable:
##                        --------------------------
##                                lnmdvalhs0
##                            (1)             (2)
## ---------------------------------------------------
## PredNPL                  -0.005          -0.001
##                          (0.027)         (0.038)
##
## ---------------------------------------------------
## Observations              483             226
## R2                       0.851           0.889
## Adjusted R2              0.823           0.836
## Residual Std. Error 0.186 (df = 407) 0.171 (df = 153)
## ===================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

In the IV (2SLS) methods, both specification lead to statistically insignificant coefficient. Hence, we observe that impact of being put on NPL list by 2000 leads to little change in value of homes from 1980 to 2000.

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.

In the above 2SLS procedure, we find that being placed on NPL by year 2000 seems to have statistically insignificant impact on housing prices. This would give interpretation that WTP or willingness to pay for hazardous waste is very small. However, we should note there could be additional factors why we get insignificant results. For example, if the clean-up has not started in many sites by year 2,000 and clean-up generally takes a long time, then not all the 'benefits' or 'willingness to pay for clean-up' may have been internalized by year 2000. In this case, running the analysis 10 years later with updated housing prices may give different result. Alternatively, clean-up process may cause negative externalities (i.e. noise or trucks passing through) and depress housing demand while in progress and perhaps after clean-up is finished, given improved amenities, supply of housing increases. In any of these cases, results could be measuring something slightly other than Willingness to Pay for CleanUp.

Appendix A: R Code

```
rm(list=ls())
knitr::opts_chunk$set(echo = F, warning = FALSE, message = FALSE)
# stargazer table type (html, latex, or text)
# Change to latex when outputting to PDF, html when outputting to html
table_type = "latex"
library(tidyverse)
library(haven)
library(stargazer)
library(ggplot2)
library(tinytex)
library(lfe) # For felm() fixed effects model
library(lmtest)
library(sandwich)
library(cobalt)
library(kableExtra)
allsite <- read_dta('allsites.dta')
allcovariates <- read_dta('allcovariates.dta')
twomiledata <- read_dta('2miledata.dta')
sitecovariates <- read_dta('sitecovariates.dta')
# In "allsite.dta", we observe that there are duplicate observations. Dropped
sum(duplicated(allsite$fips)) # There are cases of duplicate observations with same FIPs.
sum(duplicated(allsite)) # This confirms that their entire row values are
                         # duplicates and can be dropped.
# Duplicates Dropped for "allsite.dta"
allsite <- allsite[!duplicated(allsite),]

# Duplicates Dropped for "allcovarites.dta" - same procedure as above
sum(duplicated(allcovariates))
allcovariates <- allcovariates[!duplicated(allcovariates),]

# No duplicates in "twomiledata.dta"
sum(duplicated(twomiledata))
# Regress 2000 housing prices on 1(NPL site in 2000) and 1980 housing values.
reg_1a1 <- lm(lnmdvalhs0 ~ npl2000 + lnmeanhs8, data = allsite)
reg_1a1_r <- coeftest(reg_1a1, vcov = vcovHC(reg_1a1, type = "HC1"))

# Regress: Add housing characteristics as control.
HousingChar <- c("smhse8", "tothsun8", "ownocc8", "firestoveheat80", "noaircond80",
                 "nofullkitchen80", "zerofullbath80",
                "bedrms1_80occ", "bedrms2_80occ", "bedrms3_80occ", "bedrms4_80occ", "bedrms5_80occ",
                "blt2_5yrs80occ", "blt6_10yrs80occ", "blt10_20yrs80occ", "blt20_30yrs80occ", "blt30_40y
                "detach80occ", "attach80occ", "mobile80occ", "occupied80")
Formula1 <- formula(paste("lnmdvalhs0 ~ npl2000 + lnmeanhs8 + ", paste(HousingChar, collapse=" + ")))
reg_1a2 <- lm(Formula1, data = allsite)
reg_1a2_r <- coeftest(reg_1a2, vcov = vcovHC(reg_1a2, type = "HC1"))

# Regress: Add economic and demographic variables as control.
EconNDemo <- c("pop_den8", "shrblk8", "shrhsp8", "child8", "old8", "shrfor8", "ffh8",
               "hsdrop8", "no_hs_diploma8", "ba_or_better8", "unemprt8", "povrat8",
               "welfare8", "avhhin8")
Formula2 <- formula(paste("lnmdvalhs0 ~ npl2000 + lnmeanhs8 + ", paste(c(HousingChar, EconNDemo), colla
```

```
reg_1a3 <- lm(Formula2, data = allsite)
reg_1a3_r <- coeftest(reg_1a3, vcov = vcovHC(reg_1a3, type = "HC1"))

# Include State FE
Formula3 <- formula(paste("lnmdvalhs0 ~ npl2000 + lnmeanhs8 + ", paste(c(HousingChar, EconNDemo), collap
reg_1a4 <- felm(Formula3, data = allsite)

# Stargazer Table with All Four Regressions
stargazer(reg_1a1_r, reg_1a2_r, reg_1a3_r, reg_1a4,
          se = list(reg_1a1_r[,2], reg_1a2_r[,2], reg_1a3_r[,2], reg_1a4$rse),
          keep = "npl2000", type = "text", omit.stat = "f")
# Using "allcovariates.dta", compare covariates between treatment and comparison groups.
# Note: variable "old8" is not available in "allcovariates.dta"
ttest_1 <- allcovariates %>%
                mutate(npl2000_f = ifelse(npl2000 == 1, "OnList", "NotOnList")) %>%
                select(npl2000_f,
                          smhse8, tothsun8, ownocc8, firestoveheat80,
                          noaircond80, nofullkitchen80, zerofullbath80,
                          bedrms0_80occ, bedrms1_80occ, bedrms2_80occ,
                            bedrms3_80occ, bedrms4_80occ, bedrms5_80occ,
                          blt0_1yrs80occ, blt2_5yrs80occ, blt6_10yrs80occ,
                            blt10_20yrs80occ, blt20_30yrs80occ, blt30_40yrs80occ, blt40_yrs80occ,
                          detach80occ, attach80occ, mobile80occ, occupied80,
                          pop_den8, shrblk8, shrhsp8, child8, shrfor8, ffh8,
                          hsdrop8, no_hs_diploma8, ba_or_better8, unemprt8, povrat8,
                          welfare8, avhhin8) %>%
                gather(key = variable, value = value, - npl2000_f) %>%
                group_by(npl2000_f, variable) %>%
                summarise(value = list(value)) %>%
                spread(npl2000_f, value) %>%
                group_by(variable) %>%
                mutate(p_value = t.test(unlist(OnList), unlist(NotOnList))$p.value,
                       t_value = t.test(unlist(OnList), unlist(NotOnList))$statistic) %>%
                select(variable, p_value, t_value) %>%
                rename(p_val_sp1 = p_value, t_val_sp1 = t_value)

# Using "sitecovariates.dta", compare covarites between census tracts with HRS test score from 1982 abo
# Note: variable "old8" is not available in "allcovariates.dta"
ttest_2 <- sitecovariates %>%
                mutate(above = ifelse(hrs_82 > 28.5, "Above", "Below")) %>%
                select(above,
                          smhse8, tothsun8, ownocc8, firestoveheat80,
                          noaircond80, nofullkitchen80, zerofullbath80,
                          bedrms0_80occ, bedrms1_80occ, bedrms2_80occ,
                            bedrms3_80occ, bedrms4_80occ, bedrms5_80occ,
                          blt0_1yrs80occ, blt2_5yrs80occ, blt6_10yrs80occ,
                            blt10_20yrs80occ, blt20_30yrs80occ, blt30_40yrs80occ, blt40_yrs80occ,
                          detach80occ, attach80occ, mobile80occ, occupied80,
                          pop_den8, shrblk8, shrhsp8, child8, shrfor8, ffh8,
                          hsdrop8, no_hs_diploma8, ba_or_better8, unemprt8, povrat8,
                          welfare8, avhhin8) %>%
                gather(key = variable, value = value, - above) %>%
                group_by(above, variable) %>%
```

```
                summarise(value = list(value)) %>%
                spread(above, value) %>%
                group_by(variable) %>%
                mutate(p_value = t.test(unlist(Above), unlist(Below))$p.value,
                       t_value = t.test(unlist(Above), unlist(Below))$statistic) %>%
                select(variable, p_value, t_value) %>%
                rename(p_val_sp2 = p_value, t_val_sp2 = t_value)

# Using "sitecovaraites.dta", compare covarites between census tracts with HRS test score from 1982 of
ttest_3 <- sitecovariates %>%
                filter(hrs_82 >= 16.5 & hrs_82 <= 40.5) %>%
                mutate(justabove = ifelse(hrs_82 > 28.5, "JustAbove", "JustBelow")) %>%
                select(justabove,
                          smhse8, tothsun8, ownocc8, firestoveheat80,
                          noaircond80, nofullkitchen80, zerofullbath80,
                          bedrms0_80occ, bedrms1_80occ, bedrms2_80occ,
                            bedrms3_80occ, bedrms4_80occ, bedrms5_80occ,
                          blt0_1yrs80occ, blt2_5yrs80occ, blt6_10yrs80occ,
                            blt10_20yrs80occ, blt20_30yrs80occ, blt30_40yrs80occ, blt40_yrs80occ,
                          detach80occ, attach80occ, mobile80occ, occupied80,
                          pop_den8, shrblk8, shrhsp8, child8, shrfor8, ffh8,
                          hsdrop8, no_hs_diploma8, ba_or_better8, unemprt8, povrat8,
                          welfare8, avhhin8) %>%
                gather(key = variable, value = value, - justabove) %>%
                group_by(justabove, variable) %>%
                summarise(value = list(value)) %>%
                spread(justabove, value) %>%
                group_by(variable) %>%
                mutate(p_value = t.test(unlist(JustAbove), unlist(JustBelow))$p.value,
                       t_value = t.test(unlist(JustAbove), unlist(JustBelow))$statistic) %>%
                select(variable, p_value, t_value) %>%
                rename(p_val_sp3 = p_value, t_val_sp3 = t_value)

# Output t-test results
merged <- left_join(ttest_1, ttest_2, by = 'variable')
merged <- left_join(merged, ttest_3, by = 'variable')
merged %>% mutate_if(is.numeric, round, digits = 3) %>% print.data.frame()
x <- c("smhse8", "tothsun8", "ownocc8", "firestoveheat80", "noaircond80",
       "nofullkitchen80", "zerofullbath80",
       "bedrms1_80occ", "bedrms2_80occ", "bedrms3_80occ", "bedrms4_80occ", "bedrms5_80occ",
       "blt2_5yrs80occ", "blt6_10yrs80occ", "blt10_20yrs80occ", "blt20_30yrs80occ", "blt30_40yrs80occ",
       "detach80occ", "attach80occ", "mobile80occ", "occupied80",
       "pop_den8", "shrblk8", "shrhsp8", "child8", "shrfor8", "ffh8",
       "hsdrop8", "no_hs_diploma8", "ba_or_better8", "unemprt8", "povrat8",
       "welfare8", "avhhin8")

allcovariates %>%
    distinct() %>%
    select(all_of(x), npl2000) %>%
    love.plot(., treat='npl2000', data=., limits = c(-1, 1),
            stats="m")

allcovariates %>%
```

```
    distinct() %>%
    select(npl2000) %>%
    data.frame() %>%
    stargazer(omit.summary.stat = c('sd', 'p25', 'p75'), header=F)
sitecovariates %>%
    select(all_of(x), hrs_82) %>%
    mutate(treat = ifelse(hrs_82 >= 28.5, 1, 0)) %>%
    select(all_of(x), treat) %>%
    love.plot(., treat='treat', data=., limits = c(-1, 1),
            stats="m")

sitecovariates %>%
    select(all_of(x), hrs_82) %>%
    mutate(treat = ifelse(hrs_82 >= 28.5, 1, 0)) %>%
    select(treat) %>%
    data.frame() %>%
    stargazer(omit.summary.stat = c('sd', 'p25', 'p75'), header=F, type = "text")

# 16.5 and 28.5 to census tracts with sites between 28.5 and 40.5
sitecovariates %>%
    select(all_of(x), hrs_82, -npl2000) %>%
    mutate(treat = case_when(hrs_82 >= 16.5 & hrs_82 < 28.5 ~ 0,
                             hrs_82 >= 28.5 & hrs_82 < 40.5 ~ 1,
                             TRUE ~ as.numeric(NA))) %>%
    filter(!is.na(treat)) %>%
    select(all_of(x), treat) %>%
    love.plot(., treat='treat', data=., limits = c(-1, 1),
            stats="m")

sitecovariates %>%
    select(all_of(x), hrs_82, -npl2000) %>%
    mutate(treat = case_when(hrs_82 >= 16.5 & hrs_82 < 28.5 ~ 0,
                             hrs_82 >= 28.5 & hrs_82 < 40.5 ~ 1,
                             TRUE ~ as.numeric(NA))) %>%
    filter(!is.na(treat)) %>%
    select(treat) %>%
    data.frame() %>%
    stargazer(omit.summary.stat = c('sd', 'p25', 'p75'), header=F, type = "text")
#Setting bandwidth and bins
h = 1.5
bins = seq(from = 0, to = 75, by = h)
#Creating histogram with density data
hist_hrs82 <- hist(twomiledata$hrs_82, breaks = bins, freq = FALSE)
data_hist <- data.frame(mid = hist_hrs82$mids, density = hist_hrs82$density)
#Regression under
reg_under = predict(lm(density ~ mid, subset(data_hist, mid < 28.5)))
#Regression over
reg_over = predict(lm(density ~ mid, subset(data_hist, mid > 28.5)))
#Histogram with local linear regressions
plot(hist_hrs82, main = "Checking for Bunching",
     freq = FALSE,
     xlim = c(0, 80),
     xlab = "HRS Score in 1982") +
```

```
abline(v = 28.5, lty = "dashed", col = "BLUE", lwd = 3) +
lines(data_hist$mid[data_hist$mid < 28.5], reg_under) +
lines(data_hist$mid[data_hist$mid > 28.5], reg_over)
h = 1.5
cutpoints = seq(16.5, 40.5, by=1.5)
midpoints = seq(16.5+h/2, 40.5-h/2, by=1.5)
# Create counts for each bin
cut(read_dta('sitecovariates.dta')$hrs_82,
                  breaks=cutpoints,
                  include.lowest=TRUE,
                  right=FALSE,
    labels=midpoints) %>%
    summary() %>%
    data.frame() %>%
    rename(count = ".") %>%
    mutate(midpoint = rownames(.)) %>%
    filter(midpoint != "NA's") %>%
    mutate(midpoint = as.numeric(midpoint),
           treat = ifelse(midpoint > 28.5, 1, 0)) %>%
    ggplot(aes(x=midpoint, y=count, group = factor(treat))) + geom_col() +
    # Linear regressions on either side of the threshold
    geom_smooth(method=lm, se=T) +
    geom_vline(aes(xintercept=28.5), color = 'red', size=3) +
    xlab('HRS score in 1982') +
    ggtitle('Counts of HRS site scores', subtitle = 'Linear regressions using counts and bin midpoints')
    annotate(geom = "segment", x = 24.1, y = 20, xend = 28, yend = 20,
             arrow = arrow(length = unit(2, "mm"))) +
    annotate(geom = "text", x = 24, y = 20,
             label = "HRS Score Threshold (28.5)",
             hjust = "right")

names(twomiledata) <- sub("\\_nbr$", "", names(twomiledata))

# Run 1st stage regression (as written above)
twomiledata$ind_hrsabove = twomiledata$hrs_82 >= 28.5
Formula4 <- formula(paste("npl2000 ~ ind_hrsabove + ", paste(c(HousingChar, EconNDemo), collapse=" + ")

reg3a_1 <- felm(Formula4, data = twomiledata)

# Run 1st stage regression (limiting to tracts with HRS_82 between 16.5 and 40.5)

twomiledata_subset <- twomiledata[twomiledata$hrs_82 >= 16.5,]
twomiledata_subset <- twomiledata_subset[twomiledata_subset$hrs_82 <= 40.5,]

reg3a_2 <- felm(Formula4, data = twomiledata_subset)

# Output Table
stargazer(reg3a_1,reg3a_2,
          se = list(reg3a_1$rse, reg3a_2$rse),
          keep = "ind_hrsabove", type = "text", omit.stat = "f")

plot(twomiledata$hrs_82, twomiledata$npl2000,
  xlab = "HRS Score in 1982", ylab = "Whether on NPL by Year 2000")
```

```
  abline(v = 28.5, lty = 3, col = 2, lwd = 2)
plot(twomiledata$hrs_82, twomiledata$lnmeanhs8,
  xlab = "HRS Score in 1982", ylab = "Median House Value in 1980")
  abline(v = 28.5, lty = 3, col = 2, lwd = 2)
  abline(lm(lnmeanhs8 ~ hrs_82, data = twomiledata), col = "blue")

# Run 2SLS (tracts with HRS_82 below 28.5 and above)
twomiledata$PredNPL <- reg3a_1$fitted.values
Formula5 <- formula(paste("lnmdvalhs0 ~ PredNPL + ", paste(c(HousingChar, EconNDemo), collapse=" + "),
reg4_1 <- felm(Formula5, data = twomiledata)


# Run 2SLS (limiting to tracts with HRS_82 between 16.5 and 40.5)
twomiledata_subset$PredNPL <- reg3a_2$fitted.values
reg4_2 <- felm(Formula5, data = twomiledata_subset)

# Output Table
stargazer(reg4_1, reg4_2,
          se = list(reg4_1$rse, reg4_2$rse),
          keep = "PredNPL", type = "text", omit.stat = "f")
```