ARE 213 Problem Set 3

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

12/7/2020

Question 1: OLS Regressions

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?

```r
## Regress 2000 housing prices on whether the census tract had an NPL site in 2000
## Use "robust" standard errors for all regressions

# 1980 housing values as a control
lm1 <- lm(lnmdvalhs0 ~ npl2000 + lnmeanhs8, allSites)
summary(lm1, se = "white")
```

```
##
## Call:
## lm(formula = lnmdvalhs0 ~ npl2000 + lnmeanhs8, data = allSites)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0581 -0.2173 -0.0241  0.1921  2.8403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.404265   0.038359  62.678  < 2e-16 ***
## npl2000     0.040036   0.013080   3.061  0.00221 **
## lnmeanhs8   0.855746   0.003519 243.174  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4057 on 42971 degrees of freedom
## Multiple R-squared:  0.5792, Adjusted R-squared:  0.5791
## F-statistic: 2.957e+04 on 2 and 42971 DF,  p-value: < 2.2e-16
```

```r
# Add housing characteristics as controls
lm2 <- lm(lnmdvalhs0 ~ npl2000 + lnmeanhs8 + tothsun8 + ownocc8 + firestoveheat80 +
            noaircond80 + nofullkitchen80 + zerofullbath80 + bedrms0_80occ + bedrms1_80occ +
            bedrms2_80occ + bedrms3_80occ + bedrms4_80occ + blt0_1yrs80occ +
            blt2_5yrs80occ + blt6_10yrs80occ + blt10_20yrs80occ + blt20_30yrs80occ +
            blt30_40yrs80occ + detach80occ + attach80occ +
            occupied80, allSites)
summary(lm2, se = "white")
```

```
##
```

```
## Call:
## lm(formula = lnmdvalhs0 ~ npl2000 + lnmeanhs8 + tothsun8 + ownocc8 +
##     firestoveheat80 + noaircond80 + nofullkitchen80 + zerofullbath80 +
##     bedrms0_80occ + bedrms1_80occ + bedrms2_80occ + bedrms3_80occ +
##     bedrms4_80occ + blt0_1yrs80occ + blt2_5yrs80occ + blt6_10yrs80occ +
##     blt10_20yrs80occ + blt20_30yrs80occ + blt30_40yrs80occ +
##     detach80occ + attach80occ + occupied80, data = allSites)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8411 -0.1843 -0.0061  0.1883  2.8498
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.675e+00  7.066e-02  37.865  < 2e-16 ***
## npl2000           4.638e-02  1.192e-02   3.892 9.96e-05 ***
## lnmeanhs8         8.525e-01  4.117e-03 207.082  < 2e-16 ***
## tothsun8          1.078e-05  4.151e-06   2.596  0.00944 **
## ownocc8          -9.363e-05  6.349e-06 -14.747  < 2e-16 ***
## firestoveheat80   5.065e-03  2.226e-02   0.228  0.82002
## noaircond80       3.030e-01  7.117e-03  42.577  < 2e-16 ***
## nofullkitchen80  -1.552e+00  1.145e-01 -13.560  < 2e-16 ***
## zerofullbath80    6.332e-01  9.332e-02   6.786 1.17e-11 ***
## bedrms0_80occ    -8.136e-01  1.554e-01  -5.237 1.64e-07 ***
## bedrms1_80occ    -4.476e-01  5.687e-02  -7.871 3.59e-15 ***
## bedrms2_80occ    -9.906e-01  4.410e-02 -22.464  < 2e-16 ***
## bedrms3_80occ    -1.104e+00  4.319e-02 -25.570  < 2e-16 ***
## bedrms4_80occ    -6.559e-01  5.522e-02 -11.879  < 2e-16 ***
## blt0_1yrs80occ   -5.293e-02  4.042e-02  -1.309  0.19042
## blt2_5yrs80occ   -1.460e-01  2.241e-02  -6.512 7.48e-11 ***
## blt6_10yrs80occ  -8.761e-02  2.015e-02  -4.349 1.37e-05 ***
## blt10_20yrs80occ -4.340e-02  1.433e-02  -3.028  0.00247 **
## blt20_30yrs80occ  1.909e-02  1.398e-02   1.365  0.17219
## blt30_40yrs80occ -8.915e-02  2.208e-02  -4.037 5.42e-05 ***
## detach80occ      -2.392e-02  2.299e-02  -1.040  0.29821
## attach80occ       2.398e-01  2.433e-02   9.857  < 2e-16 ***
## occupied80        7.268e-01  3.582e-02  20.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3681 on 42951 degrees of freedom
## Multiple R-squared:  0.6537, Adjusted R-squared:  0.6535
## F-statistic:  3685 on 22 and 42951 DF,  p-value: < 2.2e-16
```

```r
# Add economic and demographic variables as controls
lm3 <- lm(lnmdvalhs0 ~ npl2000 + lnmeanhs8 + tothsun8 + ownocc8 + firestoveheat80 +
            noaircond80 + nofullkitchen80 + zerofullbath80 + bedrms0_80occ + bedrms1_80occ +
            bedrms2_80occ + bedrms3_80occ + bedrms4_80occ + blt0_1yrs80occ +
            blt2_5yrs80occ + blt6_10yrs80occ + blt10_20yrs80occ + blt20_30yrs80occ +
            blt30_40yrs80occ + detach80occ + attach80occ +
            occupied80 + pop_den8 + shrblk8 + shrhsp8 + child8 + old8 + shrfor8 + ffh8 + smhse8 + hsdrop8 +
            no_hs_diploma8 + ba_or_better8 + unemprt8 + povrat8 + welfare8 + avhhin8, allSites)
summary(lm3, se = "white")
```

```
##
## Call:
## lm(formula = lnmdvalhs0 ~ npl2000 + lnmeanhs8 + tothsun8 + ownocc8 +
##     firestoveheat80 + noaircond80 + nofullkitchen80 + zerofullbath80 +
##     bedrms0_80occ + bedrms1_80occ + bedrms2_80occ + bedrms3_80occ +
##     bedrms4_80occ + blt0_1yrs80occ + blt2_5yrs80occ + blt6_10yrs80occ +
```

```
##     blt10_20yrs80occ + blt20_30yrs80occ + blt30_40yrs80occ +
##     detach80occ + attach80occ + occupied80 + pop_den8 + shrblk8 +
##     shrhsp8 + child8 + old8 + shrfor8 + ffh8 + smhse8 + hsdrop8 +
##     no_hs_diploma8 + ba_or_better8 + unemprt8 + povrat8 + welfare8 +
##     avhhin8, data = allSites)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8592 -0.1767  0.0039  0.1817  2.2852
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.774e+00  7.879e-02  73.279  < 2e-16 ***
## npl2000           7.329e-02  1.055e-02   6.949 3.74e-12 ***
## lnmeanhs8         5.579e-01  4.813e-03 115.907  < 2e-16 ***
## tothsun8          1.369e-05  4.638e-06   2.953 0.003151 **
## ownocc8          -1.376e-04  7.406e-06 -18.584  < 2e-16 ***
## firestoveheat80  -1.928e-02  2.023e-02  -0.953 0.340571
## noaircond80       4.349e-01  7.031e-03  61.859  < 2e-16 ***
## nofullkitchen80  -5.867e-01  1.022e-01  -5.739 9.61e-09 ***
## zerofullbath80    5.220e-01  8.492e-02   6.147 7.96e-10 ***
## bedrms0_80occ    -7.532e-01  1.381e-01  -5.455 4.91e-08 ***
## bedrms1_80occ    -2.038e-01  5.284e-02  -3.856 0.000115 ***
## bedrms2_80occ    -4.163e-01  4.182e-02  -9.954  < 2e-16 ***
## bedrms3_80occ    -5.606e-01  4.092e-02 -13.702  < 2e-16 ***
## bedrms4_80occ    -5.013e-01  4.948e-02 -10.132  < 2e-16 ***
## blt0_1yrs80occ    1.283e-01  3.712e-02   3.456 0.000548 ***
## blt2_5yrs80occ    1.649e-01  2.309e-02   7.141 9.39e-13 ***
## blt6_10yrs80occ   1.190e-01  1.864e-02   6.385 1.73e-10 ***
## blt10_20yrs80occ  6.726e-02  1.325e-02   5.078 3.83e-07 ***
## blt20_30yrs80occ  1.629e-02  1.292e-02   1.261 0.207326
## blt30_40yrs80occ  2.729e-02  2.045e-02   1.334 0.182144
## detach80occ      -2.731e-01  2.082e-02 -13.115  < 2e-16 ***
## attach80occ      -2.636e-01  2.269e-02 -11.615  < 2e-16 ***
## occupied80        1.389e-01  3.339e-02   4.159 3.20e-05 ***
## pop_den8          7.304e-06  2.560e-07  28.535  < 2e-16 ***
## shrblk8          -1.099e-01  1.157e-02  -9.493  < 2e-16 ***
## shrhsp8          -2.836e-01  1.719e-02 -16.498  < 2e-16 ***
## child8           -5.682e-01  4.255e-02 -13.353  < 2e-16 ***
## old8             -3.318e-01  3.664e-02  -9.056  < 2e-16 ***
## shrfor8           1.219e+00  2.883e-02  42.286  < 2e-16 ***
## ffh8             -5.043e-02  2.575e-02  -1.958 0.050181 .
## smhse8            4.332e-01  1.910e-02  22.686  < 2e-16 ***
## hsdrop8           1.893e-02  1.899e-02   0.997 0.318866
## no_hs_diploma8   -3.163e-01  2.546e-02 -12.424  < 2e-16 ***
## ba_or_better8     4.719e-01  2.757e-02  17.116  < 2e-16 ***
## unemprt8         -1.217e+00  5.538e-02 -21.978  < 2e-16 ***
## povrat8          -3.725e-01  3.752e-02  -9.930  < 2e-16 ***
## welfare8          8.612e-01  4.843e-02  17.783  < 2e-16 ***
## avhhin8           1.317e-05  3.289e-07  40.038  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3244 on 42936 degrees of freedom
## Multiple R-squared:  0.7312, Adjusted R-squared:  0.731
## F-statistic:  3156 on 37 and 42936 DF,  p-value: < 2.2e-16
```

```
# Add state fixed effects
lm4 <- feols(lnmdvalhs0 ~ npl2000 + lnmeanhs8 + tothsun8 + ownocc8 + firestoveheat80 +
```

```r
            noaircond80 + nofullkitchen80 + zerofullbath80 + bedrms0_80occ + bedrms1_80occ +
            bedrms2_80occ + bedrms3_80occ + bedrms4_80occ + blt0_1yrs80occ +
            blt2_5yrs80occ + blt6_10yrs80occ + blt10_20yrs80occ + blt20_30yrs80occ +
            blt30_40yrs80occ + detach80occ + attach80occ + occupied80 +
            pop_den8 + shrblk8 + shrhsp8 + child8 + old8 + shrfor8 + ffh8 + smhse8 + hsdrop8 +
            no_hs_diploma8 + ba_or_better8 + unemprt8 + povrat8 + welfare8 + avhhin8,
            fixef = "statefips", data = allSites)
summary(lm4, se = "white")
```

```
## OLS estimation, Dep. Var.: lnmdvalhs0
## Observations: 42,974
## Fixed-effects: statefips: 51
## Standard-errors: White
##                    Estimate Std. Error    t value  Pr(>|t|)
## npl2000           0.0665050 0.008798000   7.559000  4.14e-14 ***
## lnmeanhs8         0.4963600 0.020677000  24.005000 < 2.2e-16 ***
## tothsun8          0.0000140 0.000006290   2.176700  0.029512 *
## ownocc8          -0.0001310 0.000009900 -13.262000 < 2.2e-16 ***
## firestoveheat80   0.0782240 0.032040000   2.441400  0.014633 *
## noaircond80       0.3252810 0.009481000  34.310000 < 2.2e-16 ***
## nofullkitchen80  -0.5434770 0.146254000  -3.716000  0.000203 ***
## zerofullbath80    0.5912470 0.114421000   5.167300  2.39e-07 ***
## bedrms0_80occ    -0.4724540 0.224005000  -2.109100  0.034939 *
## bedrms1_80occ    -0.1237030 0.074514000  -1.660100  0.096895 .
## bedrms2_80occ    -0.3755640 0.055512000  -6.765400  1.35e-11 ***
## bedrms3_80occ    -0.4423300 0.053825000  -8.217900 < 2.2e-16 ***
## bedrms4_80occ    -0.4756380 0.063119000  -7.535600  4.96e-14 ***
## blt0_1yrs80occ    0.0962100 0.045011000   2.137500  0.032564 *
## blt2_5yrs80occ    0.1126610 0.026267000   4.289000   1.8e-05 ***
## blt6_10yrs80occ   0.0403750 0.020822000   1.939000  0.052506 .
## blt10_20yrs80occ -0.0217200 0.014083000  -1.542300  0.123014
## blt20_30yrs80occ -0.0274560 0.013036000  -2.106200  0.035194 *
## blt30_40yrs80occ  0.0117780 0.022507000   0.523329  0.600748
## detach80occ      -0.2525420 0.022135000 -11.409000 < 2.2e-16 ***
## attach80occ      -0.1888220 0.025403000  -7.433000  1.08e-13 ***
## occupied80       -0.0261770 0.043578000  -0.600698  0.548045
## pop_den8          0.0000068 0.000000391  17.372000 < 2.2e-16 ***
## shrblk8          -0.0961310 0.012675000  -7.584600   3.4e-14 ***
## shrhsp8          -0.0892790 0.021243000  -4.202700   2.6e-05 ***
## child8           -0.3620550 0.051954000  -6.968800  3.24e-12 ***
## old8             -0.1969480 0.043514000  -4.526100  6.02e-06 ***
## shrfor8           0.5883980 0.039192000  15.013000 < 2.2e-16 ***
## ffh8             -0.1120600 0.033249000  -3.370400  0.000751 ***
## smhse8            0.3755790 0.022963000  16.356000 < 2.2e-16 ***
## hsdrop8           0.0216240 0.023180000   0.932870  0.350892
## no_hs_diploma8   -0.3557730 0.033012000 -10.777000 < 2.2e-16 ***
## ba_or_better8     0.5065340 0.034710000  14.593000 < 2.2e-16 ***
## unemprt8         -1.3813000 0.074458000 -18.551000 < 2.2e-16 ***
## povrat8          -0.0930000 0.047739000  -1.948100  0.051409 .
## welfare8          0.2943820 0.064874000   4.537800   5.7e-06 ***
## avhhin8           0.0000140 0.000000622  22.203000 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -8,338.51   Adj. R2: 0.77886
##                             R2-Within: 0.68765
```

Taking our 4th regression as an example (with all controls and state fixed effects), we can interpret the results as having a hazardous waste site listed on the NPL by 2000 is associated with a 6.65 % increase in median housing values in 2000, ceteris paribus. The coefficients on NPL 2000 status will be unbiased if the conditional independence assumption holds i.e if we have selection on observables ($(Y_i(1), Y_i(0)) \perp D_i | X_i$). The key assumption underlying a "selection on observables" design is that the treatment is as good as randomly assigned after we condition on observables. In other words, we assume that we observe *all* the factors that affect treatment assignment (whether a census tract contained a hazardous waste site that was placed on the NPL by 2000) and are correlated with the potential outcomes (2000 housing values). If there is systematic selection into treatment, we assume this selection is only a function of the observables. That is, we assume that having a hazardous waste site placed on the NPL by 2000 is uncorrelated with unobservable determinants of 2000 housing values conditional on the observables. If these assumptions hold, then a regression of 2000 housing prices on whether the census tract had an NPL site in 2000, conditioning on observables, will estimate the ATE. Thus, the coefficients on NPL 2000 status will be unbiased under these conditions.

(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?

```r
## Compare covariates between census tracts with and without a hazardous waste site listed
## on the NPL by 2000 using allcovariates.dta

# no old8 in allcovariates.dta

# Make a log mean housing prices 1980 variable
allCovariates$lnmeanhs80 <- log(allCovariates$meanhs8)

summary_dt <- transpose(allCovariates[,lapply(.SD, mean, na.rm=TRUE), .SDcols = c(59, 23, 24, 28:31,
                                                                                   40:56, 3:15, 17),
                               by = npl2000])
summary_dt <- cbind(summary_dt, transpose(allCovariates[,lapply(.SD, sd, na.rm=TRUE), .SDcols =
                                                       c(59, 23, 24, 28:31, 40:56, 3:15, 17),
                                                  by = npl2000]))
colnames(summary_dt) <- c("No NPL means", "NPL means", "No NPL sd", "NPL sd")
summary_dt <- summary_dt[2:39,]
summary_dt[, "Variable (1980 Values)" := c("Log mean housing values", "Total housing units in tract",
                           "# owner-occupied units", "% units fire stove heat",
                           "% units with no AC", "% units no full kitchen",
                           "% units with no full bath", "% own-occ units 0 bedrms",
                           "% own-occ units 1 bedrm",  "% own-occ units 2 bedrms",
                           "% own-occ units 3 bedrms",  "% own-occ units 4 bedrms",
                           "% own-occ units 5 bedrms", "% own-occ units built in last yr",
                           "% own-occ units 2-5 yrs old", "% own-occ units 6-10 yrs old",
                           "% own-occ units 10-20 yrs old", "% own-occ units 20-30 yrs old",
                           "% own-occ units 30-40 yrs old", "% own-occ units 40+ yrs old",
                           "% detached single family housing", "% attached single family housing",
                           "% mobile home single family", "% housing units occupied",
                           "Tract population density", "Share of pop black", "Share of pop Hispanic",
                           "Share of population < 18", "Share of population foreign born",
                           "Share of female headed HHs", "% pop in same house 5 yrs ago",
                           "% pop high school aged HS dropout", "% of pop > 25 with no HS diploma",
                           "% pop > 25 with BA or better", "% pop > 16 unemployed",
                           "% pop below poverty line", "% of HHs on public assistance",
                           "Average household income")]

formulas <- paste("allCovariates$", names(allCovariates)[c(59, 23, 24, 28:31, 40:56, 3:15, 17)],
                  "~ allCovariates$npl2000")
t_test <- t(sapply(formulas, function(f) {
  res <- t.test(as.formula(f))
  c(res$statistic, p.value=res$p.value)
```

```r
}))

colnames(t_test) <- c("t-stat", "p-value")

summary_dt <- cbind(summary_dt, t_test)
summary_dt[, Difference := `No NPL means` - `NPL means`]

setcolorder(summary_dt, c("Variable (1980 Values)", "No NPL means", "No NPL sd", "NPL means",
                         "NPL sd", "Difference", "t-stat", "p-value"))
```

```r
# First table for 1b

print(xtable(summary_dt, caption = "Difference in Means of Census Tracks Without Versus With a Site Listed on
```

Variable (1980 Values)	No NPL means	No NPL sd	NPL means	NPL sd	Difference	t-stat	p-value
Log mean housing values	10.88	0.55	10.81	0.46	0.07	4.61	0.00
Total housing units in tract	1347.86	690.84	1392.00	637.36	-44.13	-2.15	0.03
# owner-occupied units	800.65	464.12	907.86	463.36	-107.21	-7.19	0.00
% units fire stove heat	0.04	0.09	0.05	0.08	-0.01	-4.03	0.00
% units with no AC	0.43	0.29	0.49	0.25	-0.06	-7.84	0.00
% units no full kitchen	0.02	0.03	0.02	0.03	-0.00	-1.77	0.08
% units with no full bath	0.02	0.04	0.03	0.03	-0.00	-2.45	0.01
% own-occ units 0 bedrms	0.00	0.01	0.00	0.01	0.00	2.19	0.03
% own-occ units 1 bedrm	0.05	0.06	0.04	0.04	0.00	2.03	0.04
% own-occ units 2 bedrms	0.26	0.15	0.28	0.12	-0.01	-3.36	0.00
% own-occ units 3 bedrms	0.48	0.14	0.48	0.11	-0.00	-1.31	0.19
% own-occ units 4 bedrms	0.17	0.11	0.16	0.09	0.01	3.28	0.00
% own-occ units 5 bedrms	0.04	0.05	0.03	0.03	0.00	4.75	0.00
% own-occ units built in last yr	0.04	0.06	0.03	0.04	0.00	2.93	0.00
% own-occ units 2-5 yrs old	0.11	0.13	0.11	0.10	-0.00	-0.64	0.52
% own-occ units 6-10 yrs old	0.12	0.12	0.13	0.10	-0.01	-3.85	0.00
% own-occ units 10-20 yrs old	0.19	0.16	0.19	0.12	-0.01	-2.10	0.04
% own-occ units 20-30 yrs old	0.18	0.16	0.19	0.14	-0.01	-1.87	0.06
% own-occ units 30-40 yrs old	0.10	0.11	0.11	0.09	-0.00	-0.87	0.38
% own-occ units 40+ yrs old	0.26	0.28	0.23	0.21	0.03	4.55	0.00
% detached single family housing	0.88	0.20	0.88	0.16	0.00	0.54	0.59
% attached single family housing	0.07	0.18	0.04	0.11	0.03	9.56	0.00
% mobile home single family	0.05	0.10	0.09	0.12	-0.04	-9.33	0.00
% housing units occupied	0.93	0.06	0.94	0.04	-0.01	-5.03	0.00
Tract population density	5424.07	9479.35	1406.95	2267.74	4017.13	47.60	0.00
Share of pop black	0.12	0.24	0.09	0.19	0.03	4.03	0.00
Share of pop Hispanic	0.07	0.14	0.05	0.12	0.02	5.38	0.00
Share of population < 18	0.28	0.07	0.29	0.06	-0.01	-7.60	0.00
Share of population foreign born	0.07	0.09	0.05	0.07	0.02	6.88	0.00
Share of female headed HHs	0.19	0.14	0.16	0.12	0.03	7.31	0.00
% pop in same house 5 yrs ago	0.52	0.16	0.54	0.14	-0.03	-6.40	0.00
% pop high school aged HS dropout	0.14	0.11	0.14	0.10	-0.00	-0.54	0.59
% of pop > 25 with no HS diploma	0.31	0.17	0.34	0.15	-0.03	-5.92	0.00
% pop > 25 with BA or better	0.17	0.13	0.14	0.10	0.04	11.48	0.00
% pop > 16 unemployed	0.07	0.04	0.07	0.04	-0.00	-2.76	0.01
% pop below poverty line	0.11	0.10	0.11	0.09	0.01	2.39	0.02
% of HHs on public assistance	0.08	0.08	0.07	0.07	0.00	1.31	0.19
Average household income	21510.18	8616.42	20340.16	6348.18	1170.02	5.68	0.00

Table 1: Difference in Means of Census Tracks Without Versus With a Site Listed on the NPL by 2000

Table 1 suggests that census tracts with a hazardous waste site listed on the NPL by 2000 differ systematically from census tracts without such sites. Specifically, compared to tracts with such sites, census tracts without a site listed on the NPL by 2000 have on average higher mean housing values, fewer housing units, fewer owner-occupied units, smaller proportion of units with no AC, smaller proportion of units with no full bath, higher population density, higher share of black and hispanic populations, higher average household income, among other differences, all of which are statistically significant at at least the 5% level. These results suggest that restricting comparisons to census tracts with a site listed on the NPL by 2000 may be a better strategy.

```
## Compare covariates between census tracts with a hazardous waste site that had
## an HRS test in 1982 using sitecovariates.dta. Specifically, compare sites that
## scored above 28.5 to those that scored below 28.5

# no old8 in sitecovariates.dta

# Make a log mean housing prices 1980 variable
siteCovariates$lnmeanhs80 <- log(siteCovariates$meanhs8)

# Make a dummy for scoring <= 28.5 == 0 and scoring > 28.5 == 1
siteCovariates$HRSaboveThreshold <- 3
siteCovariates[, HRSaboveThreshold := ifelse(hrs_82 > 28.5, 1, 0)]

summary_dt_2 <- transpose(siteCovariates[,lapply(.SD, mean, na.rm=TRUE), .SDcols = c(61, 24, 25, 29:32,
                                                                                      41:57, 4:16, 18),
                                by = HRSaboveThreshold])
summary_dt_2 <- cbind(summary_dt_2, transpose(siteCovariates[,lapply(.SD, sd, na.rm=TRUE), .SDcols =
                                                              c(61, 24, 25, 29:32, 41:57, 4:16, 18),
                                              by = HRSaboveThreshold]))
colnames(summary_dt_2) <- c("Below mean", "Above mean", "Below sd", "Above sd")
summary_dt_2 <- summary_dt_2[2:39,]
summary_dt_2[, "Variable (1980 Values)" := c("Log mean housing values", "Total housing units in tract",
                           "# owner-occupied units", "% units fire stove heat",
                           "% units with no AC", "% units no full kitchen",
                           "% units with no full bath", "% own-occ units 0 bedrms",
                           "% own-occ units 1 bedrm",  "% own-occ units 2 bedrms",
                           "% own-occ units 3 bedrms",  "% own-occ units 4 bedrms",
                           "% own-occ units 5 bedrms", "% own-occ units built in last yr",
                           "% own-occ units 2-5 yrs old", "% own-occ units 6-10 yrs old",
                           "% own-occ units 10-20 yrs old", "% own-occ units 20-30 yrs old",
                           "% own-occ units 30-40 yrs old", "% own-occ units 40+ yrs old",
                           "% detached single family housing", "% attached single family housing",
                           "% mobile home single family", "% housing units occupied",
                           "Tract population density", "Share of pop black", "Share of pop Hispanic",
                           "Share of population < 18", "Share of population foreign born",
                           "Share of female headed HHs", "% pop in same house 5 yrs ago",
                           "% pop high school aged HS dropout", "% of pop > 25 with no HS diploma",
                           "% pop > 25 with BA or better", "% pop > 16 unemployed",
                           "% pop below poverty line", "% of HHs on public assistance",
                           "Average household income")]

formulas <- paste("siteCovariates$", names(siteCovariates)[c(61, 24, 25, 29:32, 41:57, 4:16, 18)],
                  "~ siteCovariates$HRSaboveThreshold")
t_test <- t(sapply(formulas, function(f) {
  res <- t.test(as.formula(f))
  c(res$statistic, p.value=res$p.value)
}))

colnames(t_test) <- c("t-stat", "p-value")

summary_dt_2 <- cbind(summary_dt_2, t_test)
summary_dt_2[, Diff := `Below mean` - `Above mean`]

setcolorder(summary_dt_2, c("Variable (1980 Values)", "Below mean", "Below sd", "Above mean",
                          "Above sd", "Diff", "t-stat", "p-value"))
```

```r
# Second table for 1b

print(xtable(summary_dt_2, caption = "Difference in Means of Census Tracks with a HRS Score Below Versus Abov
      include.rownames = FALSE, size = "small", comment = FALSE)
```

Variable (1980 Values)	Below mean	Below sd	Above mean	Above sd	Diff	t-stat	p-value
Log mean housing values	10.63	0.41	10.79	0.39	-0.15	-4.11	0.00
Total housing units in tract	1356.74	703.15	1352.81	630.37	3.93	0.06	0.95
# owner-occupied units	906.37	505.43	901.85	461.03	4.52	0.10	0.92
% units fire stove heat	0.05	0.07	0.05	0.08	0.00	0.24	0.81
% units with no AC	0.51	0.24	0.48	0.24	0.03	1.14	0.25
% units no full kitchen	0.02	0.03	0.02	0.03	0.00	0.84	0.40
% units with no full bath	0.03	0.04	0.03	0.03	0.01	1.70	0.09
% own-occ units 0 bedrms	0.00	0.01	0.00	0.01	-0.00	-0.74	0.46
% own-occ units 1 bedrm	0.05	0.05	0.04	0.05	0.00	0.50	0.62
% own-occ units 2 bedrms	0.31	0.13	0.27	0.12	0.04	3.35	0.00
% own-occ units 3 bedrms	0.48	0.12	0.49	0.11	-0.01	-0.47	0.64
% own-occ units 4 bedrms	0.13	0.07	0.16	0.09	-0.03	-4.11	0.00
% own-occ units 5 bedrms	0.03	0.02	0.03	0.03	-0.01	-2.57	0.01
% own-occ units built in last yr	0.03	0.03	0.03	0.04	-0.01	-2.04	0.04
% own-occ units 2-5 yrs old	0.09	0.10	0.10	0.10	-0.01	-1.57	0.12
% own-occ units 6-10 yrs old	0.11	0.09	0.13	0.10	-0.02	-2.38	0.02
% own-occ units 10-20 yrs old	0.18	0.12	0.20	0.12	-0.02	-2.25	0.03
% own-occ units 20-30 yrs old	0.18	0.14	0.19	0.13	-0.01	-0.64	0.52
% own-occ units 30-40 yrs old	0.11	0.09	0.10	0.08	0.01	1.88	0.06
% own-occ units 40+ yrs old	0.30	0.25	0.24	0.22	0.06	2.70	0.01
% detached single family housing	0.86	0.20	0.89	0.14	-0.03	-1.97	0.05
% attached single family housing	0.06	0.18	0.03	0.09	0.03	2.06	0.04
% mobile home single family	0.08	0.11	0.08	0.11	0.00	0.26	0.79
% housing units occupied	0.94	0.04	0.94	0.04	-0.00	-0.08	0.94
Tract population density	1670.24	3508.63	1157.46	1772.99	512.78	1.83	0.07
Share of pop black	0.11	0.23	0.07	0.16	0.04	2.10	0.04
Share of pop Hispanic	0.04	0.10	0.04	0.11	0.00	0.20	0.84
Share of population < 18	0.29	0.06	0.29	0.06	-0.00	-0.05	0.96
Share of population foreign born	0.05	0.09	0.05	0.07	0.00	0.34	0.74
Share of female headed HHs	0.19	0.15	0.16	0.12	0.03	2.39	0.02
% pop in same house 5 yrs ago	0.60	0.13	0.56	0.13	0.04	3.28	0.00
% pop high school aged HS dropout	0.14	0.09	0.13	0.10	0.01	1.19	0.24
% of pop > 25 with no HS diploma	0.41	0.15	0.34	0.14	0.06	4.68	0.00
% pop > 25 with BA or better	0.10	0.07	0.14	0.10	-0.04	-4.92	0.00
% pop > 16 unemployed	0.09	0.05	0.07	0.04	0.02	3.39	0.00
% pop below poverty line	0.11	0.10	0.10	0.08	0.01	1.61	0.11
% of HHs on public assistance	0.09	0.08	0.07	0.07	0.01	2.05	0.04
Average household income	19635.32	4942.86	20868.85	5797.29	-1233.53	-2.49	0.01

Table 2: Difference in Means of Census Tracks with a HRS Score Below Versus Above the 28.5 Threshold

The results from Table 2 suggest that among census tracts with a hazardous waste site that had an HRS test in 1982, tracts with a site that scored above the 28.5 threshold differ systematically from sites with a HRS score below the threshold. For example, census tracts with a site scoring below the threshold have on average lower mean housing values, higher share of female-headed households, higher share of population in the same house 5 years ago, higher proportion of population with no high school diploma, lower proportion of population with a college degree or better, a higher unemployment rate, a higher proportion of households on public assistance, and lower household income, among other differences, all of which are statistically significant at at least the 5% level.

```r
## Compare covariates between census tracts with a hazardous waste site that had
## an HRS test in 1982 using sitecovariates.dta. Specifically, compare sites that
## scored between 16.5 and 28.5 to those that scored between 28.5 and 40.5

# no old8 in sitecovariates.dta
```

```r
# drop observations with HRS below 16.5 or above 40.5
siteCovariates <- subset(siteCovariates, siteCovariates$hrs_82 <= 40.5)
siteCovariates <- subset(siteCovariates, siteCovariates$hrs_82 >= 16.5)

# HRSaboveThreshold from before is still the appropriate dummy since it keeps track of
# above or below 28.5 and we've removed observations outside the desired intervals

summary_dt_3 <- transpose(siteCovariates[,lapply(.SD, mean, na.rm=TRUE), .SDcols = c(61, 24, 25, 29:32,
                                                                                       41:57, 4:16, 18),
                              by = HRSaboveThreshold])
summary_dt_3 <- cbind(summary_dt_3, transpose(siteCovariates[,lapply(.SD, sd, na.rm=TRUE), .SDcols =
                                                           c(61, 24, 25, 29:32, 41:57, 4:16, 18),
                                                 by = HRSaboveThreshold]))
colnames(summary_dt_3) <- c("Below mean", "Above mean", "Below sd", "Above sd")
summary_dt_3 <- summary_dt_3[2:39,]
summary_dt_3[, "Variable (1980 Values)" := c("Log mean housing values", "Total housing units in tract",
                           "# owner-occupied units", "% units fire stove heat",
                           "% units with no AC", "% units no full kitchen",
                           "% units with no full bath", "% own-occ units 0 bedrms",
                           "% own-occ units 1 bedrm",  "% own-occ units 2 bedrms",
                           "% own-occ units 3 bedrms",  "% own-occ units 4 bedrms",
                           "% own-occ units 5 bedrms", "% own-occ units built in last yr",
                           "% own-occ units 2-5 yrs old", "% own-occ units 6-10 yrs old",
                           "% own-occ units 10-20 yrs old", "% own-occ units 20-30 yrs old",
                           "% own-occ units 30-40 yrs old", "% own-occ units 40+ yrs old",
                           "% detached single family housing", "% attached single family housing",
                           "% mobile home single family", "% housing units occupied",
                           "Tract population density", "Share of pop black", "Share of pop Hispanic",
                           "Share of population < 18", "Share of population foreign born",
                           "Share of female headed HHs", "% pop in same house 5 yrs ago",
                           "% pop high school aged HS dropout", "% of pop > 25 with no HS diploma",
                           "% pop > 25 with BA or better", "% pop > 16 unemployed",
                           "% pop below poverty line", "% of HHs on public assistance",
                           "Average household income")]

formulas <- paste("siteCovariates$", names(siteCovariates)[c(61, 24, 25, 29:32, 41:57, 4:16, 18)],
                  "~ siteCovariates$HRSaboveThreshold")
t_test <- t(sapply(formulas, function(f) {
  res <- t.test(as.formula(f))
  c(res$statistic, p.value=res$p.value)
}))

colnames(t_test) <- c("t-stat", "p-value")

summary_dt_3 <- cbind(summary_dt_3, t_test)
summary_dt_3[, Diff := `Below mean` - `Above mean`]

setcolorder(summary_dt_3, c("Variable (1980 Values)", "Below mean", "Below sd", "Above mean", "Above sd", "Di
```

```r
# Third table for 1b

print(xtable(summary_dt_3, caption = "Difference in Means of Census Tracks with a HRS Score Below Versus Abov
      include.rownames = FALSE, size = "small", comment = FALSE)
```

The results reported in Table 3 suggest that census tracts above and below the threshold become more comparable when we shorten the bandwidth. That is, when we only compare census tracts that are within 12 points of the threshold on either side, most of the differences in observable characteristics become statistically insignificant. We do find a statistically significant difference in two variables, percentage of owner-occupied units with 2 bedrooms and percentage of population

Variable (1980 Values)	Below mean	Below sd	Above mean	Above sd	Diff	t-stat	p-value
Log mean housing values	10.68	0.36	10.74	0.42	-0.06	-1.23	0.22
Total housing units in tract	1366.93	629.95	1319.33	618.75	47.60	0.56	0.58
# owner-occupied units	946.44	482.96	871.53	455.46	74.91	1.17	0.24
% units fire stove heat	0.06	0.07	0.05	0.08	0.01	0.66	0.51
% units with no AC	0.52	0.24	0.51	0.24	0.01	0.16	0.87
% units no full kitchen	0.02	0.03	0.02	0.04	0.00	0.27	0.79
% units with no full bath	0.03	0.04	0.03	0.04	0.00	0.87	0.39
% own-occ units 0 bedrms	0.00	0.01	0.00	0.01	-0.00	-0.47	0.64
% own-occ units 1 bedrm	0.05	0.04	0.05	0.06	0.00	0.13	0.90
% own-occ units 2 bedrms	0.31	0.12	0.27	0.11	0.03	2.06	0.04
% own-occ units 3 bedrms	0.47	0.12	0.48	0.10	-0.01	-0.72	0.47
% own-occ units 4 bedrms	0.14	0.07	0.16	0.09	-0.02	-1.71	0.09
% own-occ units 5 bedrms	0.03	0.03	0.04	0.04	-0.01	-1.23	0.22
% own-occ units built in last yr	0.03	0.03	0.03	0.03	-0.00	-0.05	0.96
% own-occ units 2-5 yrs old	0.10	0.10	0.10	0.09	0.00	0.01	0.99
% own-occ units 6-10 yrs old	0.12	0.09	0.13	0.09	-0.01	-0.46	0.65
% own-occ units 10-20 yrs old	0.19	0.11	0.20	0.10	-0.01	-0.40	0.69
% own-occ units 20-30 yrs old	0.19	0.13	0.19	0.12	-0.00	-0.05	0.96
% own-occ units 30-40 yrs old	0.11	0.09	0.10	0.08	0.01	0.71	0.48
% own-occ units 40+ yrs old	0.26	0.21	0.25	0.21	0.00	0.13	0.90
% detached single family housing	0.85	0.17	0.89	0.14	-0.04	-1.62	0.11
% attached single family housing	0.05	0.16	0.03	0.10	0.02	1.04	0.30
% mobile home single family	0.09	0.11	0.08	0.11	0.02	1.07	0.28
% housing units occupied	0.94	0.04	0.94	0.04	0.00	0.01	0.99
Tract population density	1360.90	3088.42	1151.05	2047.17	209.85	0.57	0.57
Share of pop black	0.08	0.21	0.08	0.18	-0.00	-0.09	0.93
Share of pop Hispanic	0.03	0.06	0.03	0.08	0.00	0.09	0.93
Share of population < 18	0.29	0.07	0.29	0.06	-0.00	-0.57	0.57
Share of population foreign born	0.04	0.05	0.04	0.04	0.00	0.27	0.79
Share of female headed HHs	0.16	0.10	0.17	0.12	-0.00	-0.17	0.86
% pop in same house 5 yrs ago	0.59	0.12	0.57	0.13	0.02	1.17	0.24
% pop high school aged HS dropout	0.15	0.10	0.13	0.09	0.01	1.03	0.30
% of pop > 25 with no HS diploma	0.39	0.13	0.35	0.14	0.03	1.89	0.06
% pop > 25 with BA or better	0.11	0.08	0.13	0.10	-0.03	-2.11	0.04
% pop > 16 unemployed	0.08	0.04	0.07	0.04	0.00	0.34	0.73
% pop below poverty line	0.11	0.09	0.11	0.09	-0.00	-0.36	0.72
% of HHs on public assistance	0.08	0.06	0.08	0.07	0.01	0.56	0.58
Average household income	19812.21	4496.38	20300.79	6026.31	-488.58	-0.70	0.49

Table 3: Difference in Means of Census Tracks with a HRS Score Below Versus Above the 28.5 Threshold (bandwidth = 12 points)

with a college degree, which is expected when we are testing balance across 38 variables at the 5% level.

Question 2: Regression Discontinuity Design

(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 below "facts" impact the appropriateness of these assumptions?

In order for regression discontinuity to be a valid research design, we need to assume that the potential outcomes $Y_i(0)$ and $Y_i(1)$ (housing prices) are smooth functions of the running variable X_i (the HRS score) as it crosses the threshold c (28.5). In other words, $E[Y_i(0)|X_i = x]$ and $E[Y_i(1)|X_i = x]$ are continuous in x. If there is imperfect compliance, that is if the probability of treatment increases, but by less than 100 pp, when the running variable crosses the threshold, then we need to use a fuzzy RD design. In this case, we need to make an additional monotonicity assumption that $D_i(x^*)$ is non-increasing in x^* at $x^* = c$, that is we need to assume there are no "defiers."

Importantly, our first assumption is violated if there is manipulation based on the HRS score. In other words, if individuals understand the assignment mechanism and can manipulate the HRS score to place a census block just above (or below) the threshold, then there is selection into treatment so census tracts just above and below the threshold are no longer comparable. Thus we need to assume that individuals cannot game the assignment mechanism in order for this to be a valid research design. Relatedly, we also need to assume that covariates are smooth at the threshold, that is that covariates are balanced above and below the threshold. If this is not true, then we have selection into treatment and observations just above and below the threshold are again not comparable.

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

This fact makes it more likely that our assumptions hold, because the threshold was not selected based on specific site characteristics, which would have potentially made covariates imbalanced across the threshold. For example, if instead the "28.5" cutoff was selected because a HRS rating of 28.5 or higher is especially (disproportionately) dangerous for human health, then our first assumption will no longer hold because houses close to sites above this threshold may benefit disproportionately from treatment.

```
(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.
```

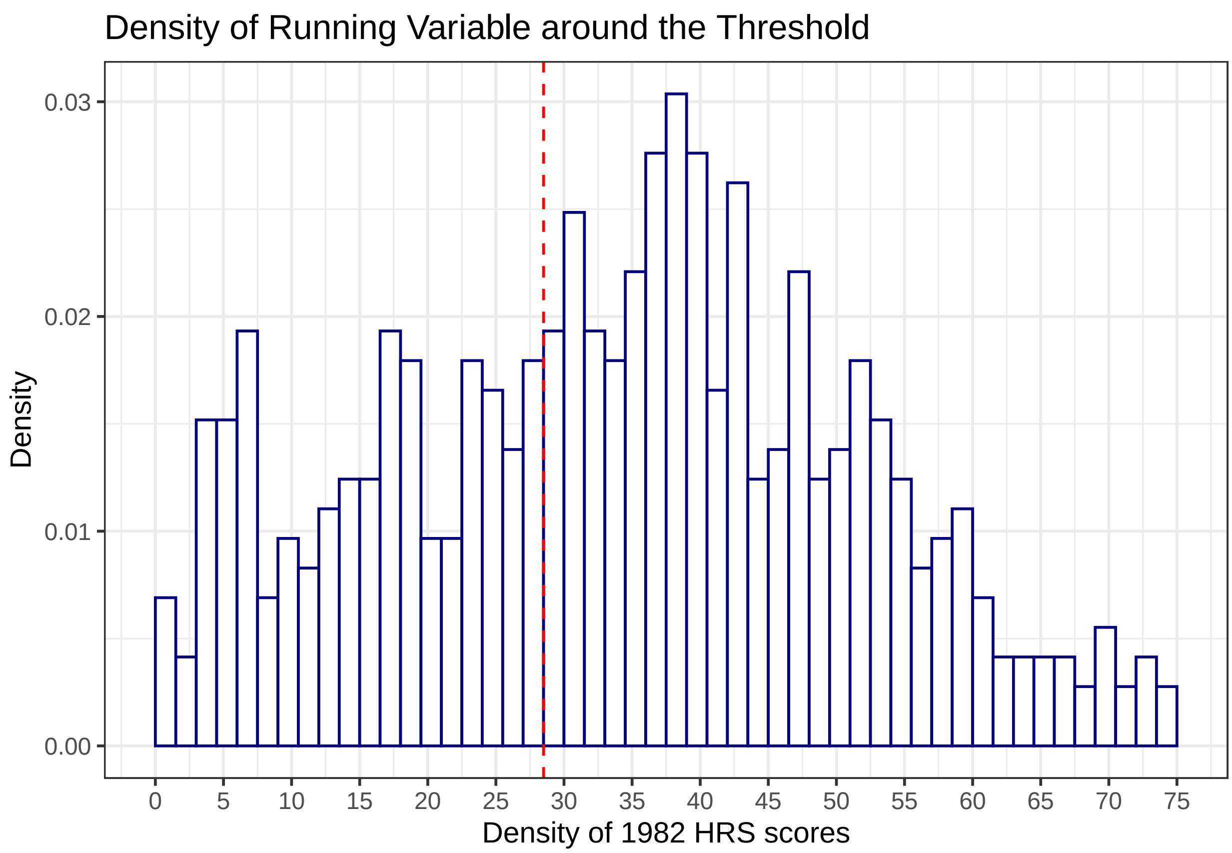
This fact makes it more likely that our assumptions hold. In particular, if none of the individuals involved knew the cutoff threshold score, it is less likely they were able to manipulate the test results to make certain census tracts be above (or below) the threshold. Even if individuals had an incentive to cheat, without knowing the assignment mechanism they would not have been able to game the system effectively.

```
(iii) EPA documentation emphasizes that the HRS test is an imperfect scoring measure
```

Whether this fact violates our assumptions depends on the type of error associated with the HRS test. If this is classical measurement error then it should not affect our assumptions. However, if the error is correlated with our covariates or with our outcome variable (housing prices) then this would violate our first assumption.

(b) Create a histogram of the distribution 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?

```r
## histogram of the density of 1982 HRS scores
ggplot(data, aes(x = hrs_82)) +
  geom_histogram(aes(y = ..density..), binwidth = 1.5, boundary = 0, closed = "left", col = "navy", fill = "w
  geom_vline(xintercept = 28.5, linetype = "dashed", color = "red") +
  theme_gray() +
  scale_x_continuous(breaks = seq(0,75,5)) +
  xlab("Density of 1982 HRS scores") +
  ylab("Density") +
  ggtitle("Density of Running Variable around the Threshold") +
  theme_bw()
```




```
## Run local linear regressions on either side of threshold, using the midpoints of the bins as the data

range(data$hrs_82)# between 0 and 74.16
```

```
## [1]  0.00 74.16
```

```
h = 1.5 #set bandwidth
bins = seq(from = 0, to = 75, by = h) # set cutoffs for bins
length(bins)
```

```
## [1] 51
```

```
# returns the bin index for each observation
data$hrs_82_bin <- cut(data$hrs_82, breaks = bins, right = FALSE)

# calculate the midpoint of each bin
bins.midpoint = (bins[-1] + bins[-(length(bins))])/2

# assign a bin midpoint to each observation
data$hrs_82_binmid = bins.midpoint[ data$hrs_82_bin ]

# generate a variable of the density of the midpoint by counting up number of observations per bin, divide by
hrs_82_bins_dist = table(data$hrs_82_binmid)/(dim(data)[1]*h)
hrs_82_bins_dist <- as.numeric(hrs_82_bins_dist)

# regression fitted on data below the cutoff
below_lm <- lm(hrs_82_bins_dist ~ bins.midpoint, data.frame(hrs_82_bins_dist, bins.midpoint)[1:19,])
summary(below_lm)
```

```
##
## Call:
## lm(formula = hrs_82_bins_dist ~ bins.midpoint, data = data.frame(hrs_82_bins_dist,
##     bins.midpoint)[1:19, ])
##
```

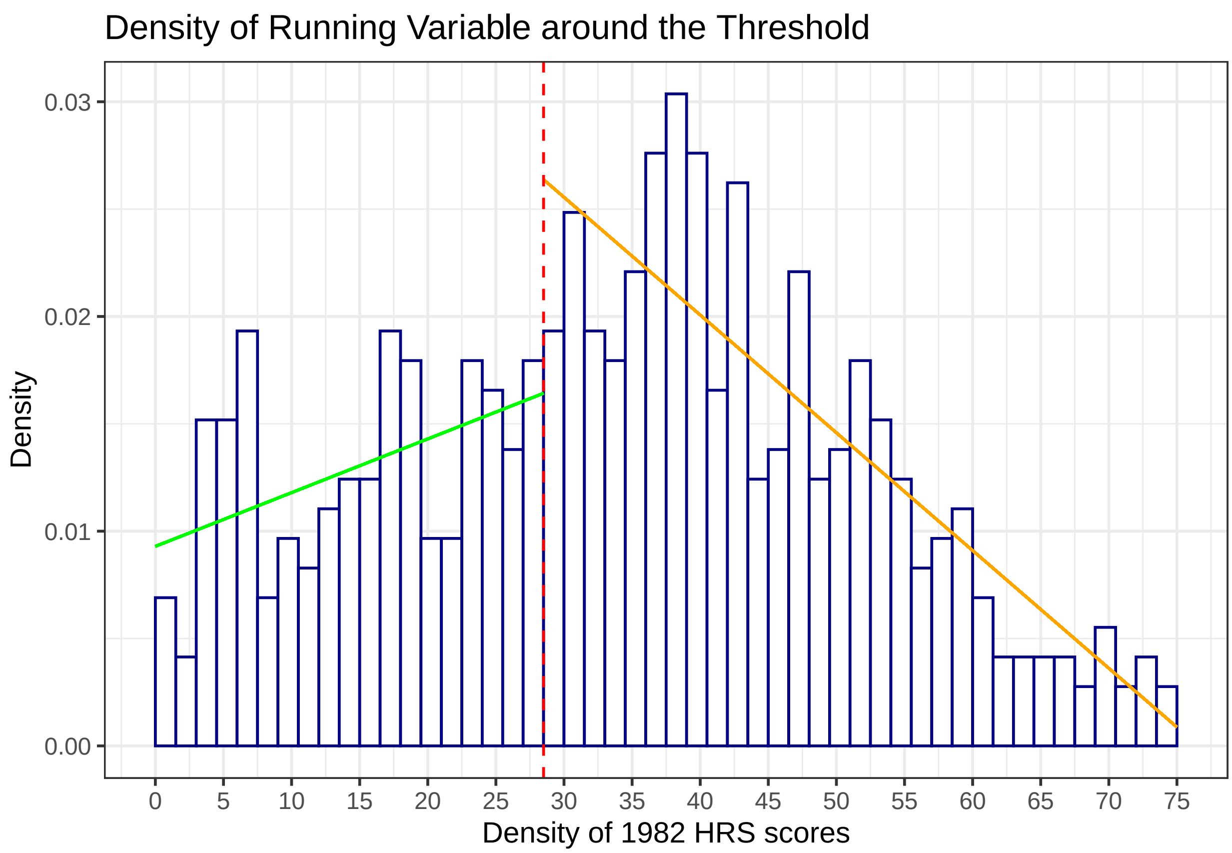
```
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0057148 -0.0032025 -0.0008112  0.0033962  0.0083421
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0092926  0.0019568   4.749 0.000186 ***
## bins.midpoint 0.0002502  0.0001190   2.103 0.050626 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00426 on 17 degrees of freedom
## Multiple R-squared:  0.2065, Adjusted R-squared:  0.1598
## F-statistic: 4.424 on 1 and 17 DF,  p-value: 0.05063
```

```
# regression fitted on data above the cutoff
above_lm <- lm(hrs_82_bins_dist ~ bins.midpoint, data.frame(hrs_82_bins_dist, bins.midpoint)[20:50,])
summary(above_lm)
```

```
##
## Call:
## lm(formula = hrs_82_bins_dist ~ bins.midpoint, data = data.frame(hrs_82_bins_dist,
##     bins.midpoint)[20:50, ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0066397 -0.0028613 -0.0006445  0.0019001  0.0093379
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.200e-02  2.984e-03  14.076 1.71e-14 ***
## bins.midpoint -5.484e-04  5.582e-05  -9.825 9.84e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00417 on 29 degrees of freedom
## Multiple R-squared:  0.769,  Adjusted R-squared:  0.761
## F-statistic: 96.52 on 1 and 29 DF,  p-value: 9.844e-11
```

```
# intercepts and slopes from fitted regression lines
ib <- below_lm[["coefficients"]][1]
sb <- below_lm[["coefficients"]][2]
ia <- above_lm[["coefficients"]][1]
sa <- above_lm[["coefficients"]][2]

# add regression lines to histogram
ggplot(data, aes(x = hrs_82)) +
  geom_histogram(aes(y = ..density..), binwidth = 1.5, boundary = 0, closed = "left", col = "navy", fill = "w
  geom_segment(aes(x = 0, xend = 28.5, y = ib, yend = ib+sb*28.5), color = "green") +
  geom_segment(aes(x = 28.5, xend = 75, y = ia + sa*28.5, yend = ia + sa*75), color = "orange") +
  geom_vline(xintercept = 28.5, linetype = "dashed", color = "red") +
  theme_gray() +
  scale_x_continuous(breaks = seq(0,75,5)) +
  xlab("Density of 1982 HRS scores") +
  ylab("Density") +
  ggtitle("Density of Running Variable around the Threshold") +
  theme_bw()
```




```r
## Repeat above but with truncated distribution, 16.5 <= HRS <= 40.5

h = 1.5 #set bandwidth
bins = seq(from = 16.5, to = 40.5, by = h) # set cutoffs for bins
length(bins)
```

```
## [1] 17
```

```r
# returns the bin index for each observation
dataT <- data[hrs_82 >= 16.5 & hrs_82 <= 40.5,]
dataT$hrs_82_bin <- cut(dataT$hrs_82, breaks = bins, right = FALSE)

# calculate the midpoint of each bin
bins.midpoint = (bins[-1] + bins[-(length(bins))])/2

# assign a bin midpoint to each observation
dataT$hrs_82_binmid = bins.midpoint[ dataT$hrs_82_bin ]

# generate a variable of the density of the midpoint by counting up number of observations per bin, divide by
hrs_82_bins_dist = table(dataT$hrs_82_binmid)/(dim(dataT)[1]*h)
hrs_82_bins_dist <- as.numeric(hrs_82_bins_dist)

# regression fitted on data below the cutoff
below_lm <- lm(hrs_82_bins_dist ~ bins.midpoint, data.frame(hrs_82_bins_dist, bins.midpoint)[1:8,])
summary(below_lm)
```

```
##
## Call:
## lm(formula = hrs_82_bins_dist ~ bins.midpoint, data = data.frame(hrs_82_bins_dist,
##     bins.midpoint)[1:8, ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

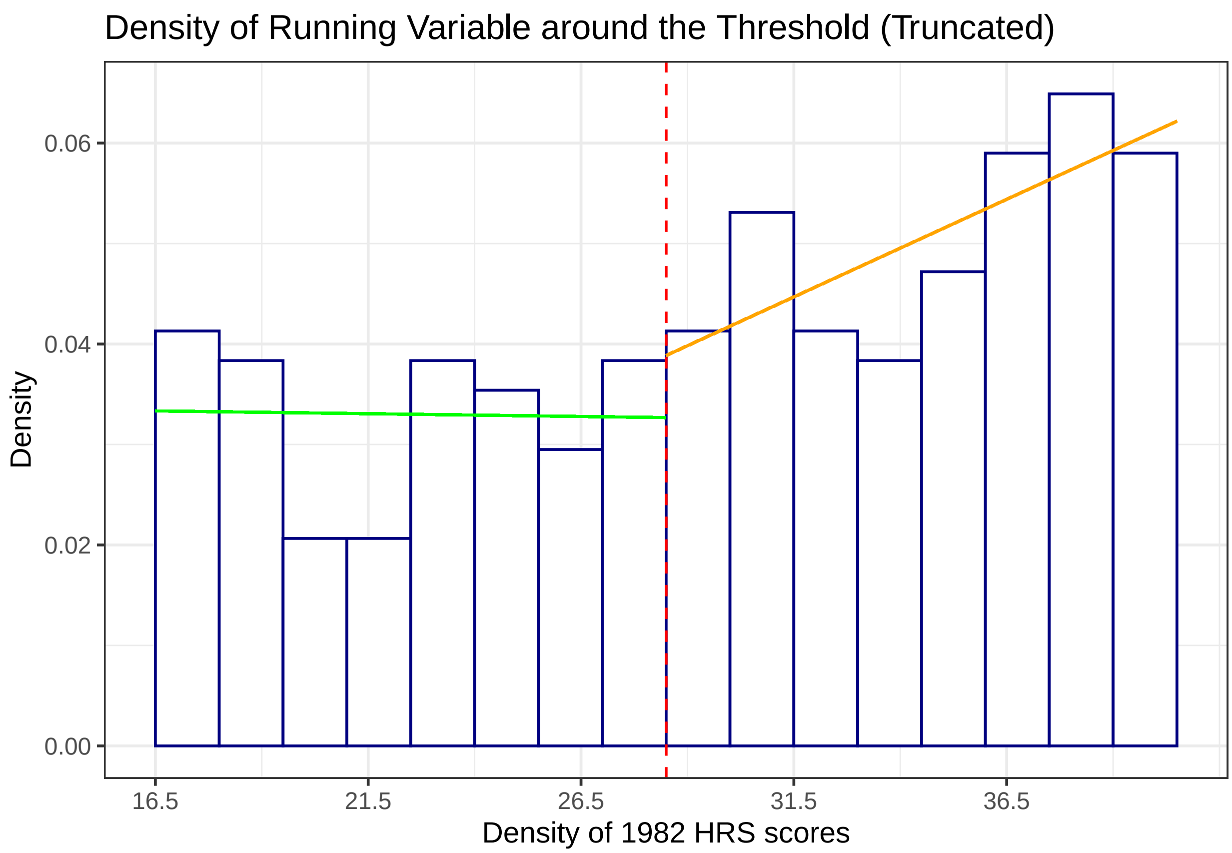
```
## -0.012221 -0.005470  0.004039  0.005575  0.008358
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.334e-02  2.089e-02   1.596    0.162
## bins.midpoint -2.341e-05  9.176e-04  -0.026    0.980
##
## Residual standard error: 0.00892 on 6 degrees of freedom
## Multiple R-squared:  0.0001085,  Adjusted R-squared:  -0.1665
## F-statistic: 0.0006509 on 1 and 6 DF,  p-value: 0.9805
```

```
# regression fitted on data above the cutoff
above_lm <- lm(hrs_82_bins_dist ~ bins.midpoint, data.frame(hrs_82_bins_dist, bins.midpoint)[9:16,])
summary(above_lm)
```

```
##
## Call:
## lm(formula = hrs_82_bins_dist ~ bins.midpoint, data = data.frame(hrs_82_bins_dist,
##     bins.midpoint)[9:16, ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0107108 -0.0047935 -0.0003687  0.0048550  0.0098680
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0165227  0.0264311  -0.625   0.5549
## bins.midpoint  0.0019432  0.0007623   2.549   0.0436 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007411 on 6 degrees of freedom
## Multiple R-squared:  0.5199, Adjusted R-squared:  0.4399
## F-statistic: 6.497 on 1 and 6 DF,  p-value: 0.04355
```

```
# intercepts and slopes from fitted regression lines
ib <- below_lm[["coefficients"]][1]
sb <- below_lm[["coefficients"]][2]
ia <- above_lm[["coefficients"]][1]
sa <- above_lm[["coefficients"]][2]

# add regression lines to histogram
ggplot(dataT, aes(x = hrs_82)) +
  geom_histogram(aes(y = ..density..), binwidth = 1.5, boundary = 0, closed = "left", col = "navy", fill = "w
  geom_segment(aes(x = 16.5, xend = 28.5, y = ib, yend = ib+sb*28.5), color = "green") +
  geom_segment(aes(x = 28.5, xend = 40.5, y = ia + sa*28.5, yend = ia + sa*40.5), color = "orange") +
  geom_vline(xintercept = 28.5, linetype = "dashed", color = "red") +
  theme_gray() +
  scale_x_continuous(breaks = seq(16.5,40.5,5)) +
  xlab("Density of 1982 HRS scores") +
  ylab("Density") +
  ggtitle("Density of Running Variable around the Threshold (Truncated)") +
  theme_bw()
```



From our histogram, it does not appear visually that there is a discontinuity in the distribution of 1982 HRS scores around the 28.5 threshold. We further test this by running local linear regressions of the density of each bin by the midpoints of each bin on either side of the threshold and add the fitted regression lines to our histogram. The local regression lines appear discontinuous at the threshold, likely because we have low density in high HRS scores driving results. When we restrict our sample to census tracts with sites with a HRS score closer to the threshold [16.5, 40.5], as we do in Questions 3 and 4 below, the local regression lines appear less discontinuous at the threshold. Especially in this localized sample, there does not appear to be a discontinuity in the density of the running variable at the 28.5 threshold.

Question 3: First Stage of RD Design

(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.

We can write the first stage as

$$NPL_i = \delta_0 + \delta_1 1(HRS_i \geq 28.5) + \gamma X_i + \nu_i$$

where the instrument for NPL status is whether HRS is above 28.5 and we control for covariates, including state fixed effects. Now we estimate this first stage regression, including the full set of controls used in the fourth regression in question 1.

```
first_stage <- feols(npl2000 ~ above_28pt5 + lnmeanhs8_nbr + tothsun8_nbr + ownocc8_nbr + firestoveheat80_nbr
            noaircond80_nbr + nofullkitchen80_nbr + zerofullbath80_nbr + bedrms0_80occ_nbr + bedrms1_80occ_nb
            bedrms2_80occ_nbr + bedrms3_80occ_nbr + bedrms4_80occ_nbr + blt0_1yrs80occ_nbr +
            blt2_5yrs80occ_nbr + blt6_10yrs80occ_nbr + blt10_20yrs80occ_nbr + blt20_30yrs80occ_nbr +
            blt30_40yrs80occ_nbr + detach80occ_nbr + attach80occ_nbr +
            occupied80_nbr + pop_den8_nbr + shrblk8_nbr + shrhsp8_nbr + child8_nbr + old8_nbr + shrfor8_nbr +
            smhse8_nbr + hsdrop8_nbr + no_hs_diploma8_nbr + ba_or_better8_nbr + unemprt8_nbr + povrat8_nbr +
```

```
           fixef = "statefips", data = data)
summary(first_stage, se = "white")
```

```
## OLS estimation, Dep. Var.: npl2000
## Observations: 483
## Fixed-effects: statefips: 40
## Standard-errors: White
##                         Estimate Std. Error   t value  Pr(>|t|)
## above_28pt5           0.799242000 0.03391900 23.563000 < 2.2e-16 ***
## lnmeanhs8_nbr        -0.022472000 0.05374000 -0.418160  0.676051
## tothsun8_nbr          0.000002530 0.00000477  0.531400  0.595432
## ownocc8_nbr          -0.000000155 0.00000746 -0.020828  0.983393
## firestoveheat80_nbr  -0.222990000 0.26359500 -0.845957  0.398075
## noaircond80_nbr       0.182903000 0.09475700  1.930200  0.054274 .
## nofullkitchen80_nbr  -0.025830000 1.75770000 -0.014695  0.988283
## zerofullbath80_nbr   -1.068500000 1.26090000 -0.847389  0.397278
## bedrms0_80occ_nbr    -2.401200000 4.12490000 -0.582127  0.560804
## bedrms1_80occ_nbr    -0.252010000 0.77130000 -0.326735  0.744037
## bedrms2_80occ_nbr     0.218320000 0.61608100  0.354369  0.723247
## bedrms3_80occ_nbr    -0.068620000 0.62917400 -0.109063  0.913206
## bedrms4_80occ_nbr    -0.410309000 0.74468300 -0.550985  0.581947
## blt0_1yrs80occ_nbr   -0.684437000 0.71178800 -0.961574  0.336836
## blt2_5yrs80occ_nbr    0.415568000 0.31423300  1.322500  0.186752
## blt6_10yrs80occ_nbr  -0.198064000 0.24605900 -0.804946  0.421322
## blt10_20yrs80occ_nbr  0.160013000 0.18922000  0.845646  0.398248
## blt20_30yrs80occ_nbr -0.026743000 0.14787100 -0.180851  0.856575
## blt30_40yrs80occ_nbr -0.529519000 0.31351200 -1.689000   0.09199 .
## detach80occ_nbr       0.033457000 0.21559600  0.155186  0.876752
## attach80occ_nbr      -0.022757000 0.23788000 -0.095667  0.923833
## occupied80_nbr       -0.644046000 0.49839500 -1.292200   0.19701
## pop_den8_nbr         -0.000008660 0.00000851 -1.017600   0.30946
## shrblk8_nbr           0.152107000 0.18739200  0.811709  0.417434
## shrhsp8_nbr           0.513898000 0.25383100  2.024600  0.043566 *
## child8_nbr           -0.339603000 0.69757100 -0.486836  0.626637
## old8_nbr             -0.479199000 0.46300600 -1.035000  0.301298
## shrfor8_nbr          -0.299393000 0.33802900 -0.885702  0.376302
## ffh8_nbr             -0.070799000 0.24289200 -0.291485  0.770829
## smhse8_nbr            0.421846000 0.26640600  1.583500  0.114092
## hsdrop8_nbr          -0.104992000 0.19436600 -0.540175  0.589372
## no_hs_diploma8_nbr   -0.155690000 0.35401500 -0.439783  0.660328
## ba_or_better8_nbr     0.404511000 0.43620200  0.927348  0.354297
## unemprt8_nbr          0.286641000 0.56945600  0.503358  0.614985
## povrat8_nbr           0.039807000 0.63627800  0.062562  0.950146
## welfare8_nbr         -0.197550000 0.63897400 -0.309168  0.757353
## avhhin8_nbr          -0.000001320 0.00000671 -0.197153  0.843806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: 70.95   Adj. R2: 0.75888
##                       R2-Within: 0.74784
```

As expected, having HRS above 28.5 is strongly predictive of NPL status. Now we rerun our first stage but focusing on census tracts with HRS between 16.5 and 40.5, which had shown the most balance across covariates in question 1b.

```
data_narrow <- data[data$hrs_82 >= 16.5 & data$hrs_82 <= 40.5,]
first_stage <- feols(npl2000 ~ above_28pt5 + lnmeanhs8_nbr + tothsun8_nbr + ownocc8_nbr + firestoveheat80_nbr
            noaircond80_nbr + nofullkitchen80_nbr + zerofullbath80_nbr + bedrms0_80occ_nbr + bedrms1_80occ_nb
            bedrms2_80occ_nbr + bedrms3_80occ_nbr + bedrms4_80occ_nbr + blt0_1yrs80occ_nbr +
            blt2_5yrs80occ_nbr + blt6_10yrs80occ_nbr + blt10_20yrs80occ_nbr + blt20_30yrs80occ_nbr +
            blt30_40yrs80occ_nbr + detach80occ_nbr + attach80occ_nbr +
```

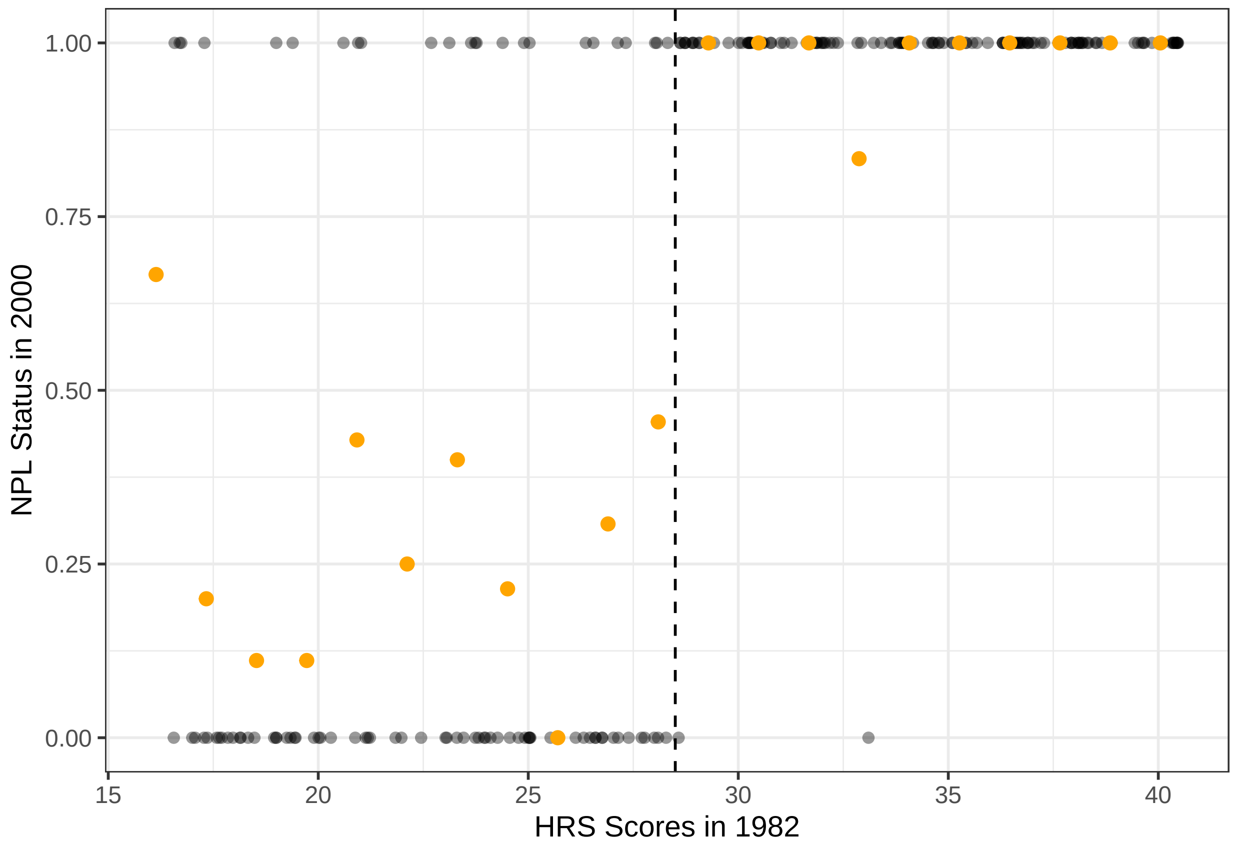
```r
            occupied80_nbr + pop_den8_nbr + shrblk8_nbr + shrhsp8_nbr + child8_nbr + old8_nbr + shrfor8_nbr +
            smhse8_nbr + hsdrop8_nbr + no_hs_diploma8_nbr + ba_or_better8_nbr + unemprt8_nbr + povrat8_nbr +
            fixef = "statefips", data = data_narrow)
summary(first_stage, se = "white")
```

```
## OLS estimation, Dep. Var.: npl2000
## Observations: 226
## Fixed-effects: statefips: 37
## Standard-errors: White
##                          Estimate Std. Error   t value  Pr(>|t|)
## above_28pt5            0.70335100   0.054072 13.008000 < 2.2e-16 ***
## lnmeanhs8_nbr          0.00856300   0.095589  0.089584  0.928736
## tothsun8_nbr           0.00000266   0.000013  0.201077  0.840907
## ownocc8_nbr            0.00000401   0.000017  0.241146  0.809767
## firestoveheat80_nbr    0.37970700   0.674109  0.563272  0.574079
## noaircond80_nbr        0.15015300   0.181166  0.828818  0.408507
## nofullkitchen80_nbr   -1.73710000   3.319600 -0.523287  0.601537
## zerofullbath80_nbr    -1.53610000   3.142800 -0.488755  0.625719
## bedrms0_80occ_nbr    -21.03000000  11.491000 -1.830200  0.069182 .
## bedrms1_80occ_nbr     -0.33123400   1.811800 -0.182822  0.855181
## bedrms2_80occ_nbr      0.45345900   1.351900  0.335433  0.737761
## bedrms3_80occ_nbr      0.16374500   1.327000  0.123392   0.90196
## bedrms4_80occ_nbr     -0.97337200   1.457500 -0.667824  0.505258
## blt0_1yrs80occ_nbr    -2.50460000   2.134900 -1.173200  0.242549
## blt2_5yrs80occ_nbr     1.85650000   0.760232  2.442000  0.015753 *
## blt6_10yrs80occ_nbr    0.41844000   0.674372  0.620487  0.535866
## blt10_20yrs80occ_nbr   0.18850700   0.467252  0.403438  0.687193
## blt20_30yrs80occ_nbr  -0.33737500   0.340570 -0.990619  0.323446
## blt30_40yrs80occ_nbr  -0.31974600   0.744782 -0.429314  0.668302
## detach80occ_nbr        0.76561400   0.587476  1.303200  0.194468
## attach80occ_nbr        0.31156800   0.655466  0.475338  0.635228
## occupied80_nbr        -0.72358400   1.338100 -0.540759  0.589465
## pop_den8_nbr          -0.00000787   0.000026 -0.306622  0.759551
## shrblk8_nbr           -0.23777900   0.407008 -0.584212  0.559944
## shrhsp8_nbr            1.19510000   0.911465  1.311200  0.191783
## child8_nbr            -1.45220000   1.271400 -1.142200  0.255147
## old8_nbr              -0.78453200   0.939741 -0.834838  0.405119
## shrfor8_nbr           -1.41270000   1.114600 -1.267400  0.206935
## ffh8_nbr               1.38140000   0.902952  1.529900  0.128121
## smhse8_nbr             1.11830000   0.537556  2.080400   0.03917 *
## hsdrop8_nbr           -0.54773900   0.531728 -1.030100  0.304594
## no_hs_diploma8_nbr     0.27481900   0.702317  0.391303  0.696121
## ba_or_better8_nbr      0.86839100   0.802171  1.082600  0.280722
## unemprt8_nbr           1.83260000   1.098600  1.668100  0.097351 .
## povrat8_nbr            0.29481400   1.442500  0.204373  0.838336
## welfare8_nbr           0.07375500   1.756100  0.041999  0.966555
## avhhin8_nbr            0.00001700   0.000016  1.063100  0.289402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -5.5503   Adj. R2: 0.56354
##                         R2-Within: 0.65361
```

Here again the threshold is strongly predictve of NPL status. This suggests that we have a strong first stage and that the indicator for HRS greater than 28.5 is a relevant instrument for NPL status in 2000.

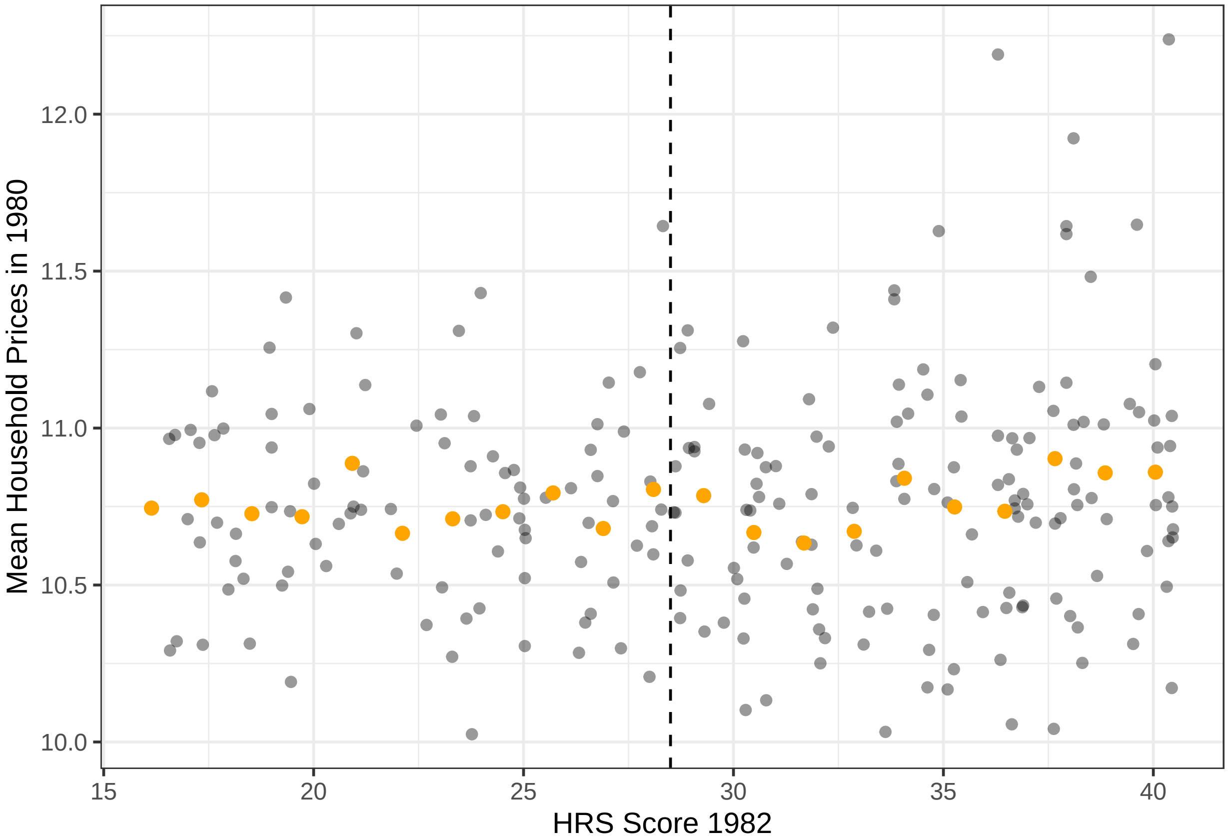
(b) Create a graph plotting the the 1982 HRS score against whether a site is listed on the NPL by year 2000 (NPL on the y-axis, HRS on the x -axis). Briefly explain and interpret this graph.

```r
(ggplot(data_narrow, aes(x=data_narrow$hrs_82,y=data_narrow$npl2000)) +
  geom_point(alpha = 0.4) +
  stat_summary_bin(fun='mean', bins=20,
                   color='orange', size=2, geom='point')) +
  geom_vline(xintercept=28.5, linetype="dashed") +
  ylab("NPL Status in 2000") +
  xlab("HRS Scores in 1982") +
  theme_bw()
```



Here the yellow dots are the binned means and the black dots are the observed values. With and HRS score below 28.5, there is still a reasonable change (around 25%) that the site will be added to the NPL. With HRS above 28.5, it is almost guaranteed (the graph shows one exception). This graph shows that the research design is *not* a sharp RD because of the probability of being treated does not go from 0 to 1 at the threshold. Instead, we have a fuzzy RD.

```r
(ggplot(data_narrow, aes(x=data_narrow$hrs_82,y=data_narrow$lnmeanhs8_nbr)) +
  geom_point(alpha = 0.4) +
  stat_summary_bin(fun='mean', bins=20,
                   color='orange', size=2, geom='point')) +
  geom_vline(xintercept=28.5, linetype="dashed") +
  ylab("Mean Household Prices in 1980") +
  xlab("HRS Score 1982") +
  theme_bw()
```



There aren't any obvious differences in this range of HRS values. Specifically, we would be concerned if there were a noticeable difference in household prices around the HRS threshold, suggesting that there was manipulation/selection to be on one side of the threshold or another. We do not see this so we believe the RD design is valid.

Question 4: Second Stage of RD Design

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.

The 2nd stage equation is

$$P_i = \beta_0 + \beta_1 N\hat{P}L_i + \psi X_i + \varepsilon_i$$

where P_i is the logged median housing price in 2000, $N\hat{P}L_i$ are the fitted values from the first stage, and the set of covariates is the same as in the first stage. The two assumptions for valid IV estimation are: 1) $Cov(z_i, d_i) \neq 0$ (relevance of the instrument) which our first stage showed we satisfy, and 2) $Cov(z_i, \epsilon_i)$ (exogeneity) which the facts and discussion in question 2 suggest we satisfy. We estimate the 2SLS regression for the full dataset and for a subset of census tracts with HRS scores between 16.5 and 40.5. Note, we could instead estimate the reduced-form regression

$$P_i = \pi_0 + \pi_1 1(HRS_i \geq 28.5) + \mu X_i + u_i$$

and then divide $\hat{pi}_1$ by our first stage coefficient $\hat{\delta}_1$. This will be equivalent to the 2SLS coefficent $\hat{\beta}_1$ and for ease of calculating standard errors, we will do 2SLS here using the package iv_robust.

```
iv_reg_full <- iv_robust(lnmdvalhs0_nbr~ npl2000 + lnmeanhs8_nbr + tothsun8_nbr + ownocc8_nbr + firestoveheat
           noaircond80_nbr + nofullkitchen80_nbr + zerofullbath80_nbr + bedrms0_80occ_nbr + bedrms1_80occ_nb
           bedrms2_80occ_nbr + bedrms3_80occ_nbr + bedrms4_80occ_nbr + blt0_1yrs80occ_nbr +
           blt2_5yrs80occ_nbr + blt6_10yrs80occ_nbr + blt10_20yrs80occ_nbr + blt20_30yrs80occ_nbr +
           blt30_40yrs80occ_nbr + detach80occ_nbr + attach80occ_nbr +
           occupied80_nbr + pop_den8_nbr + shrblk8_nbr + shrhsp8_nbr + child8_nbr + old8_nbr + shrfor8_nbr +
```

```r
            smhse8_nbr + hsdrop8_nbr + no_hs_diploma8_nbr + ba_or_better8_nbr + unemprt8_nbr + povrat8_nbr +
            welfare8_nbr + avhhin8_nbr | . - npl2000 + above_28pt5,
            data = data, fixed_effects = ~ statefips, se_type = "HC1")
summary(iv_reg_full, digits=4)
```

```
##
## Call:
## iv_robust(formula = lnmdvalhs0_nbr ~ npl2000 + lnmeanhs8_nbr +
##     tothsun8_nbr + ownocc8_nbr + firestoveheat80_nbr + noaircond80_nbr +
##     nofullkitchen80_nbr + zerofullbath80_nbr + bedrms0_80occ_nbr +
##     bedrms1_80occ_nbr + bedrms2_80occ_nbr + bedrms3_80occ_nbr +
##     bedrms4_80occ_nbr + blt0_1yrs80occ_nbr + blt2_5yrs80occ_nbr +
##
## Standard error type:  HC1
##
## Coefficients:
##                       Estimate Std. Error  t value  Pr(>|t|)   CI Lower
## npl2000             -1.734e-02  1.957e-02 -0.88601 3.761e-01 -5.582e-02
## lnmeanhs8_nbr        4.621e-01  9.115e-02  5.06997 6.060e-07  2.830e-01
## tothsun8_nbr         9.863e-06  4.634e-06  2.12828 3.392e-02  7.529e-07
## ownocc8_nbr         -2.178e-05  5.903e-06 -3.68898 2.558e-04 -3.338e-05
## firestoveheat80_nbr  2.386e-01  2.271e-01  1.05051 2.941e-01 -2.079e-01
## noaircond80_nbr      6.056e-02  7.811e-02  0.77535 4.386e-01 -9.298e-02
## nofullkitchen80_nbr -2.169e+00  1.167e+00 -1.85880 6.378e-02 -4.464e+00
## zerofullbath80_nbr   2.011e+00  1.080e+00  1.86209 6.331e-02 -1.120e-01
## bedrms0_80occ_nbr   -3.856e+00  4.645e+00 -0.83008 4.070e-01 -1.299e+01
## bedrms1_80occ_nbr    9.086e-02  7.162e-01  0.12686 8.991e-01 -1.317e+00
## bedrms2_80occ_nbr   -1.000e+00  6.016e-01 -1.66219 9.725e-02 -2.183e+00
## bedrms3_80occ_nbr   -7.442e-01  5.751e-01 -1.29412 1.964e-01 -1.875e+00
## bedrms4_80occ_nbr   -1.340e+00  6.782e-01 -1.97642 4.878e-02 -2.674e+00
## blt0_1yrs80occ_nbr  -1.253e-01  6.915e-01 -0.18117 8.563e-01 -1.485e+00
## blt2_5yrs80occ_nbr  -2.804e-01  2.527e-01 -1.10980 2.677e-01 -7.771e-01
## blt6_10yrs80occ_nbr  4.087e-01  2.165e-01  1.88747 5.981e-02 -1.697e-02
## blt10_20yrs80occ_nbr -3.052e-01  1.423e-01 -2.14472 3.257e-02 -5.849e-01
## blt20_30yrs80occ_nbr -1.221e-01  1.495e-01 -0.81692 4.145e-01 -4.159e-01
## blt30_40yrs80occ_nbr  5.821e-03  2.384e-01  0.02442 9.805e-01 -4.628e-01
## detach80occ_nbr     -4.706e-01  2.119e-01 -2.22087 2.691e-02 -8.871e-01
## attach80occ_nbr     -6.815e-01  2.423e-01 -2.81252 5.154e-03 -1.158e+00
## occupied80_nbr       5.281e-01  4.635e-01  1.13944 2.552e-01 -3.830e-01
## pop_den8_nbr         4.419e-06  7.948e-06  0.55595 5.785e-01 -1.121e-05
## shrblk8_nbr         -2.788e-01  1.458e-01 -1.91231 5.654e-02 -5.654e-01
## shrhsp8_nbr         -6.640e-01  2.096e-01 -3.16715 1.656e-03 -1.076e+00
## child8_nbr          -2.182e-01  5.164e-01 -0.42249 6.729e-01 -1.233e+00
## old8_nbr            -5.459e-01  5.179e-01 -1.05412 2.925e-01 -1.564e+00
## shrfor8_nbr          9.300e-01  3.261e-01  2.85178 4.570e-03  2.889e-01
## ffh8_nbr             6.182e-01  2.631e-01  2.34910 1.930e-02  1.009e-01
## smhse8_nbr           2.902e-01  2.102e-01  1.38089 1.681e-01 -1.229e-01
## hsdrop8_nbr          7.964e-02  2.114e-01  0.37681 7.065e-01 -3.358e-01
## no_hs_diploma8_nbr   1.218e-01  2.462e-01  0.49485 6.210e-01 -3.621e-01
## ba_or_better8_nbr    5.865e-01  4.180e-01  1.40309 1.614e-01 -2.352e-01
## unemprt8_nbr        -1.428e+00  5.118e-01 -2.79031 5.514e-03 -2.434e+00
## povrat8_nbr          4.593e-01  4.424e-01  1.03815 2.998e-01 -4.104e-01
## welfare8_nbr        -3.601e-01  5.361e-01 -0.67179 5.021e-01 -1.414e+00
## avhhin8_nbr          4.236e-05  6.674e-06  6.34627 5.902e-10  2.924e-05
##                       CI Upper  DF
## npl2000              2.113e-02 406
## lnmeanhs8_nbr        6.413e-01 406
## tothsun8_nbr         1.897e-05 406
## ownocc8_nbr         -1.017e-05 406
```

```
## firestoveheat80_nbr   6.851e-01 406
## noaircond80_nbr       2.141e-01 406
## nofullkitchen80_nbr   1.249e-01 406
## zerofullbath80_nbr    4.134e+00 406
## bedrms0_80occ_nbr     5.275e+00 406
## bedrms1_80occ_nbr     1.499e+00 406
## bedrms2_80occ_nbr     1.827e-01 406
## bedrms3_80occ_nbr     3.863e-01 406
## bedrms4_80occ_nbr    -7.185e-03 406
## blt0_1yrs80occ_nbr    1.234e+00 406
## blt2_5yrs80occ_nbr    2.163e-01 406
## blt6_10yrs80occ_nbr   8.343e-01 406
## blt10_20yrs80occ_nbr -2.546e-02 406
## blt20_30yrs80occ_nbr  1.717e-01 406
## blt30_40yrs80occ_nbr  4.745e-01 406
## detach80occ_nbr      -5.404e-02 406
## attach80occ_nbr      -2.051e-01 406
## occupied80_nbr        1.439e+00 406
## pop_den8_nbr          2.004e-05 406
## shrblk8_nbr           7.801e-03 406
## shrhsp8_nbr          -2.518e-01 406
## child8_nbr            7.969e-01 406
## old8_nbr              4.721e-01 406
## shrfor8_nbr           1.571e+00 406
## ffh8_nbr              1.135e+00 406
## smhse8_nbr            7.034e-01 406
## hsdrop8_nbr           4.951e-01 406
## no_hs_diploma8_nbr    6.058e-01 406
## ba_or_better8_nbr     1.408e+00 406
## unemprt8_nbr         -4.220e-01 406
## povrat8_nbr           1.329e+00 406
## welfare8_nbr          6.937e-01 406
## avhhin8_nbr           5.548e-05 406
##
## Multiple R-squared:  0.8826 ,    Adjusted R-squared:  0.8606
## Multiple R-squared (proj. model):  0.8007 ,  Adjusted R-squared (proj. model):  0.7634
## F-statistic (proj. model): 49.04 on 37 and 406 DF,  p-value: < 2.2e-16
```

```r
iv_reg_narrow <- iv_robust(lnmdvalhs0_nbr~ npl2000 + lnmeanhs8_nbr + tothsun8_nbr + ownocc8_nbr + firestovehe
            noaircond80_nbr + nofullkitchen80_nbr + zerofullbath80_nbr + bedrms0_80occ_nbr + bedrms1_80occ_nb
            bedrms2_80occ_nbr + bedrms3_80occ_nbr + bedrms4_80occ_nbr + blt0_1yrs80occ_nbr +
            blt2_5yrs80occ_nbr + blt6_10yrs80occ_nbr + blt10_20yrs80occ_nbr + blt20_30yrs80occ_nbr +
            blt30_40yrs80occ_nbr + detach80occ_nbr + attach80occ_nbr +
            occupied80_nbr + pop_den8_nbr + shrblk8_nbr + shrhsp8_nbr + child8_nbr + old8_nbr + shrfor8_nbr +
            smhse8_nbr + hsdrop8_nbr + no_hs_diploma8_nbr + ba_or_better8_nbr + unemprt8_nbr + povrat8_nbr +
            avhhin8_nbr | . - npl2000 + above_28pt5,
            data = data_narrow, fixed_effects = ~ statefips, se_type = "HC1")
summary(iv_reg_narrow, digits = 4)
```

```
##
## Call:
## iv_robust(formula = lnmdvalhs0_nbr ~ npl2000 + lnmeanhs8_nbr +
##     tothsun8_nbr + ownocc8_nbr + firestoveheat80_nbr + noaircond80_nbr +
##     nofullkitchen80_nbr + zerofullbath80_nbr + bedrms0_80occ_nbr +
##     bedrms1_80occ_nbr + bedrms2_80occ_nbr + bedrms3_80occ_nbr +
##     bedrms4_80occ_nbr + blt0_1yrs80occ_nbr + blt2_5yrs80occ_nbr +
##
## Standard error type:  HC1
##
## Coefficients:
```

```
##                         Estimate Std. Error  t value  Pr(>|t|)   CI Lower
## npl2000               -3.843e-02  2.560e-02 -1.50103 1.354e-01 -8.901e-02
## lnmeanhs8_nbr          3.386e-01  8.841e-02  3.82986 1.870e-04  1.639e-01
## tothsun8_nbr           1.204e-05  6.999e-06  1.72066 8.735e-02 -1.785e-06
## ownocc8_nbr           -2.184e-05  8.825e-06 -2.47477 1.443e-02 -3.927e-05
## firestoveheat80_nbr   -1.344e-01  3.369e-01 -0.39882 6.906e-01 -7.999e-01
## noaircond80_nbr       -1.450e-01  1.059e-01 -1.36972 1.728e-01 -3.543e-01
## nofullkitchen80_nbr    2.498e-01  1.725e+00  0.14484 8.850e-01 -3.157e+00
## zerofullbath80_nbr     1.029e+00  1.730e+00  0.59474 5.529e-01 -2.389e+00
## bedrms0_80occ_nbr     -1.190e+01  7.191e+00 -1.65542 9.990e-02 -2.611e+01
## bedrms1_80occ_nbr      1.297e-01  8.823e-01  0.14699 8.833e-01 -1.613e+00
## bedrms2_80occ_nbr     -1.669e+00  9.006e-01 -1.85295 6.583e-02 -3.448e+00
## bedrms3_80occ_nbr     -1.233e+00  7.824e-01 -1.57644 1.170e-01 -2.779e+00
## bedrms4_80occ_nbr     -1.897e+00  8.947e-01 -2.12012 3.562e-02 -3.665e+00
## blt0_1yrs80occ_nbr     5.606e-01  9.916e-01  0.56537 5.727e-01 -1.399e+00
## blt2_5yrs80occ_nbr    -6.735e-01  4.491e-01 -1.49977 1.357e-01 -1.561e+00
## blt6_10yrs80occ_nbr    6.118e-01  3.421e-01  1.78855 7.568e-02 -6.402e-02
## blt10_20yrs80occ_nbr -4.428e-01  2.397e-01 -1.84744 6.663e-02 -9.163e-01
## blt20_30yrs80occ_nbr -2.298e-01  2.048e-01 -1.12211 2.636e-01 -6.344e-01
## blt30_40yrs80occ_nbr -1.917e-02  3.996e-01 -0.04796 9.618e-01 -8.087e-01
## detach80occ_nbr       -4.930e-01  3.890e-01 -1.26715 2.070e-01 -1.262e+00
## attach80occ_nbr       -7.437e-01  4.568e-01 -1.62824 1.055e-01 -1.646e+00
## occupied80_nbr         2.671e-01  7.530e-01  0.35463 7.234e-01 -1.221e+00
## pop_den8_nbr          -2.435e-06  1.230e-05 -0.19787 8.434e-01 -2.674e-05
## shrblk8_nbr           -1.261e-01  1.984e-01 -0.63558 5.260e-01 -5.181e-01
## shrhsp8_nbr            5.123e-01  4.295e-01  1.19298 2.347e-01 -3.361e-01
## child8_nbr            -4.119e-02  6.067e-01 -0.06789 9.460e-01 -1.240e+00
## old8_nbr              -6.364e-01  5.297e-01 -1.20149 2.314e-01 -1.683e+00
## shrfor8_nbr            2.221e-01  6.502e-01  0.34165 7.331e-01 -1.063e+00
## ffh8_nbr               1.074e-01  4.963e-01  0.21638 8.290e-01 -8.731e-01
## smhse8_nbr            -9.436e-02  2.799e-01 -0.33709 7.365e-01 -6.474e-01
## hsdrop8_nbr           -3.208e-01  2.674e-01 -1.19977 2.321e-01 -8.490e-01
## no_hs_diploma8_nbr     3.155e-01  3.593e-01  0.87826 3.812e-01 -3.943e-01
## ba_or_better8_nbr     -7.022e-02  4.051e-01 -0.17335 8.626e-01 -8.705e-01
## unemprt8_nbr          -2.014e+00  6.344e-01 -3.17542 1.812e-03 -3.268e+00
## povrat8_nbr            1.131e+00  6.375e-01  1.77363 7.813e-02 -1.288e-01
## welfare8_nbr          -2.095e+00  8.505e-01 -2.46358 1.487e-02 -3.776e+00
## avhhin8_nbr            4.463e-05  7.476e-06  5.96939 1.615e-08  2.986e-05
##                         CI Upper  DF
## npl2000                1.215e-02 152
## lnmeanhs8_nbr          5.133e-01 152
## tothsun8_nbr           2.587e-05 152
## ownocc8_nbr           -4.404e-06 152
## firestoveheat80_nbr    5.312e-01 152
## noaircond80_nbr        6.417e-02 152
## nofullkitchen80_nbr    3.657e+00 152
## zerofullbath80_nbr     4.446e+00 152
## bedrms0_80occ_nbr      2.303e+00 152
## bedrms1_80occ_nbr      1.873e+00 152
## bedrms2_80occ_nbr      1.105e-01 152
## bedrms3_80occ_nbr      3.124e-01 152
## bedrms4_80occ_nbr     -1.292e-01 152
## blt0_1yrs80occ_nbr     2.520e+00 152
## blt2_5yrs80occ_nbr     2.137e-01 152
## blt6_10yrs80occ_nbr    1.288e+00 152
## blt10_20yrs80occ_nbr  3.074e-02 152
## blt20_30yrs80occ_nbr  1.748e-01 152
## blt30_40yrs80occ_nbr  7.704e-01 152
## detach80occ_nbr        2.757e-01 152
```

```
## attach80occ_nbr       1.587e-01 152
## occupied80_nbr        1.755e+00 152
## pop_den8_nbr          2.188e-05 152
## shrblk8_nbr           2.659e-01 152
## shrhsp8_nbr           1.361e+00 152
## child8_nbr            1.157e+00 152
## old8_nbr              4.101e-01 152
## shrfor8_nbr           1.507e+00 152
## ffh8_nbr              1.088e+00 152
## smhse8_nbr            4.587e-01 152
## hsdrop8_nbr           2.075e-01 152
## no_hs_diploma8_nbr    1.025e+00 152
## ba_or_better8_nbr     7.301e-01 152
## unemprt8_nbr         -7.611e-01 152
## povrat8_nbr           2.390e+00 152
## welfare8_nbr         -4.150e-01 152
## avhhin8_nbr           5.940e-05 152
##
## Multiple R-squared:  0.9087 ,    Adjusted R-squared:  0.8649
## Multiple R-squared (proj. model):  0.8409 ,  Adjusted R-squared (proj. model):  0.7645
## F-statistic (proj. model): 25.07 on 37 and 152 DF,  p-value: < 2.2e-16
```

Unlike our results in question 1, both of our 2SLS estimations yield a coefficient on NPL status in 2000 that is not statistically different from zero. Since the balance tables in question 1 suggested that NPL status may be endogenous, it was important to instead use the plausibly exogenous HRS score threshold as an instrument for NPL status. The results fromt this estimation suggest that the OLS regression was biased.# Question 5: Conclusion

Question 5: Conclusion

In question 1, we ran 4 different OLS specifications with different sets of covariates and found that NPL status in 2000 had a statistically significant effect of between 4 and 7% on median housing values in 2000. As discussed in question 1a, these estimates would be unbiased if the conditional independence assumption held. However, Table 1 showed that there are systematic differences between census tracts with an NPL designation and those without, which may suggest that there are also a number of unobserved differences between the two tracts. Therefore, we would be concerned that NPL status is not randomly assigned conditional on observables and our OLS regressions may be biased. Tables 2 and 3 then looked at the difference in covariates between census tracts above and below the HRS score cut-off. These census tracts were more balanced, specifically in the smaller bandwidth shown in Table 3, which suggests that census tracts with sites with a HRS score just above and below the threshold are more comparable. In question 2, we evaluate the validity of an RD design. We conclude that based on the facts about how HRS score is assigned, our density plot, and our local linear regressions, that RD is a valid research design with HRS score as our running variable, as there is no discontinuity in the density of the running variable at the threshold. In question 3, our regressions show that HRS score is a strong instrument for NPL status, verifying that we should use a fuzzy RD research design. Finally, we run 2SLS in question 4, with the HRS threshold as an instrument for NPL status. The coefficient on NPL status is not statistically different from zero so we cannot reject the null that there is no effect on housing prices from being designated NPL. This is very different from the regression results in question 1 which suggested a statistically significant positive effect of NPL status. In fact, the coefficients in both 2SLS regressions were negative and the confidence intervals in the smaller bandwidth regression show that we cannot rule out an effect of NPL status ranging from -9% to 1.2%. It could be plausible that the effect would be negative if NPL status provides previously unknown information to homeowners and home buyers about the environmental quality near a house. The difference in results between question 1 and question 4 show the importance of research design and using an exogenous treatment.