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?

```
## 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
```

```
# 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 + bedrms5_80occ + blt0_1yrs80occ +
            blt2_5yrs80occ + blt6_10yrs80occ + blt10_20yrs80occ + blt20_30yrs80occ +
            blt30_40yrs80occ + blt40_yrs80occ + detach80occ + attach80occ + mobile80occ +
            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 + bedrms5_80occ + blt0_1yrs80occ + blt2_5yrs80occ +
##     blt6_10yrs80occ + blt10_20yrs80occ + blt20_30yrs80occ + blt30_40yrs80occ +
##     blt40_yrs80occ + detach80occ + attach80occ + mobile80occ +
##     occupied80, data = allSites)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8449 -0.1842 -0.0061  0.1882  2.8495
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.732e+05  1.529e+05   1.132  0.25753
## npl2000           4.640e-02  1.192e-02   3.894 9.89e-05 ***
## lnmeanhs8         8.525e-01  4.117e-03 207.081  < 2e-16 ***
## tothsun8          1.074e-05  4.151e-06   2.587  0.00967 **
## ownocc8          -9.366e-05  6.350e-06 -14.750  < 2e-16 ***
## firestoveheat80   4.906e-03  2.226e-02   0.220  0.82559
## noaircond80       3.030e-01  7.117e-03  42.577  < 2e-16 ***
## nofullkitchen80  -1.552e+00  1.145e-01 -13.560  < 2e-16 ***
## zerofullbath80    6.337e-01  9.333e-02   6.790 1.14e-11 ***
## bedrms0_80occ    -5.112e+04  1.069e+05  -0.478  0.63236
## bedrms1_80occ    -5.112e+04  1.069e+05  -0.478  0.63237
## bedrms2_80occ    -5.112e+04  1.069e+05  -0.478  0.63236
## bedrms3_80occ    -5.112e+04  1.069e+05  -0.478  0.63236
## bedrms4_80occ    -5.112e+04  1.069e+05  -0.478  0.63236
## bedrms5_80occ    -5.112e+04  1.069e+05  -0.478  0.63237
## blt0_1yrs80occ   -5.317e-02  4.043e-02  -1.315  0.18846
## blt2_5yrs80occ   -1.462e-01  2.242e-02  -6.520 7.12e-11 ***
## blt6_10yrs80occ  -8.781e-02  2.015e-02  -4.358 1.31e-05 ***
## blt10_20yrs80occ -4.348e-02  1.433e-02  -3.034  0.00242 **
## blt20_30yrs80occ  1.888e-02  1.398e-02   1.350  0.17706
## blt30_40yrs80occ -8.940e-02  2.208e-02  -4.048 5.17e-05 ***
## blt40_yrs80occ            NA         NA      NA       NA
## detach80occ      -1.220e+05  1.096e+05  -1.114  0.26546
## attach80occ      -1.220e+05  1.096e+05  -1.114  0.26546
## mobile80occ      -1.220e+05  1.096e+05  -1.114  0.26546
## occupied80        7.268e-01  3.582e-02  20.289  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3681 on 42949 degrees of freedom
## Multiple R-squared:  0.6537, Adjusted R-squared:  0.6535
## F-statistic:  3378 on 24 and 42949 DF,  p-value: < 2.2e-16
```

```
# 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 + 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 + 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 + 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 + old8 +
##     shrfor8 + ffh8 + smhse8 + hsdrop8 + no_hs_diploma8 + ba_or_better8 +
##     unemprt8 + povrat8 + welfare8 + avhhin8, data = allSites)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8582 -0.1767  0.0039  0.1821  2.2858
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.594e+05  1.348e+05   1.924 0.054301 .
## npl2000           7.332e-02  1.055e-02   6.951 3.67e-12 ***
## lnmeanhs8         5.579e-01  4.813e-03 115.911  < 2e-16 ***
## tothsun8          1.360e-05  4.638e-06   2.933 0.003357 **
## ownocc8          -1.376e-04  7.406e-06 -18.581  < 2e-16 ***
## firestoveheat80  -1.946e-02  2.023e-02  -0.962 0.336055
## noaircond80       4.349e-01  7.030e-03  61.861  < 2e-16 ***
## nofullkitchen80  -5.866e-01  1.022e-01  -5.737 9.71e-09 ***
## zerofullbath80    5.225e-01  8.492e-02   6.153 7.67e-10 ***
## bedrms0_80occ    -9.331e+04  9.417e+04  -0.991 0.321796
## bedrms1_80occ    -9.331e+04  9.417e+04  -0.991 0.321799
## bedrms2_80occ    -9.331e+04  9.417e+04  -0.991 0.321798
## bedrms3_80occ    -9.331e+04  9.417e+04  -0.991 0.321797
## bedrms4_80occ    -9.331e+04  9.417e+04  -0.991 0.321797
## bedrms5_80occ    -9.330e+04  9.417e+04  -0.991 0.321800
## blt0_1yrs80occ    1.278e-01  3.712e-02   3.443 0.000576 ***
## blt2_5yrs80occ    1.643e-01  2.309e-02   7.115 1.14e-12 ***
## blt6_10yrs80occ   1.187e-01  1.864e-02   6.369 1.92e-10 ***
## blt10_20yrs80occ  6.708e-02  1.325e-02   5.064 4.12e-07 ***
## blt20_30yrs80occ  1.595e-02  1.292e-02   1.234 0.217108
## blt30_40yrs80occ  2.685e-02  2.046e-02   1.313 0.189325
## blt40_yrs80occ           NA         NA      NA       NA
## detach80occ      -1.661e+05  9.658e+04  -1.720 0.085526 .
## attach80occ      -1.661e+05  9.658e+04  -1.720 0.085526 .
## mobile80occ      -1.661e+05  9.658e+04  -1.720 0.085526 .
## occupied80        1.390e-01  3.339e-02   4.162 3.17e-05 ***
## pop_den8          7.307e-06  2.560e-07  28.547  < 2e-16 ***
## shrblk8          -1.099e-01  1.157e-02  -9.493  < 2e-16 ***
## shrhsp8          -2.835e-01  1.719e-02 -16.495  < 2e-16 ***
## child8           -5.682e-01  4.255e-02 -13.355  < 2e-16 ***
## old8             -3.318e-01  3.664e-02  -9.054  < 2e-16 ***
## shrfor8           1.219e+00  2.883e-02  42.284  < 2e-16 ***
## ffh8             -5.051e-02  2.575e-02  -1.962 0.049817 *
## smhse8            4.327e-01  1.910e-02  22.657  < 2e-16 ***
## hsdrop8           1.866e-02  1.899e-02   0.983 0.325844
## no_hs_diploma8   -3.164e-01  2.546e-02 -12.425  < 2e-16 ***
## ba_or_better8     4.715e-01  2.757e-02  17.099  < 2e-16 ***
## unemprt8         -1.218e+00  5.538e-02 -21.990  < 2e-16 ***
## povrat8          -3.723e-01  3.752e-02  -9.925  < 2e-16 ***
## welfare8          8.618e-01  4.843e-02  17.795  < 2e-16 ***
## avhhin8           1.318e-05  3.289e-07  40.063  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3244 on 42934 degrees of freedom
## Multiple R-squared:  0.7312, Adjusted R-squared:  0.731
## F-statistic:  2995 on 39 and 42934 DF,  p-value: < 2.2e-16
```

```
# Add state fixed effects
lm4 <- feols(lnmdvalhs0 ~ npl2000 + lnmeanhs8 + 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 + old8 + shrfor8 + ffh8 + smhse8 + hsdrop8 +
             no_hs_diploma8 + ba_or_better8 + unemprt8 + povrat8 + welfare8 + avhhin8,
             fixef = "statefips", data = allSites)
```

```
## Variables 'bedrms5_80occ', 'blt40_yrs80occ' and 'mobile80occ' have been removed because of collinearity (s
```

```
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 ***
## ... 3 variables were removed because of collinearity (bedrms5_80occ, blt40_yrs80occ and mobile80occ)
## ---
## 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?

```
## 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)
}))

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"))
```

```
# First table for 1b

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

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("HRS Below means", "HRS Above means", "HRS Below sd", "HRS 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",
```

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

```
                        "% 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[, Difference := `HRS Below means` - `HRS Above means`]

setcolorder(summary_dt_2, c("Variable (1980 Values)", "HRS Below means", "HRS Below sd", "HRS Above means",
                            "HRS Above sd", "Difference", "t-stat", "p-value"))
```

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

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.

```
## 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

# 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("HRS Below means", "HRS Above means", "HRS Below sd", "HRS 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[, Difference := `HRS Below means` - `HRS Above means`]

setcolorder(summary_dt_3, c("Variable (1980 Values)", "HRS Below means", "HRS Below sd", "HRS Above means", "
```

```
# 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)
```

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

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

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 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.
```

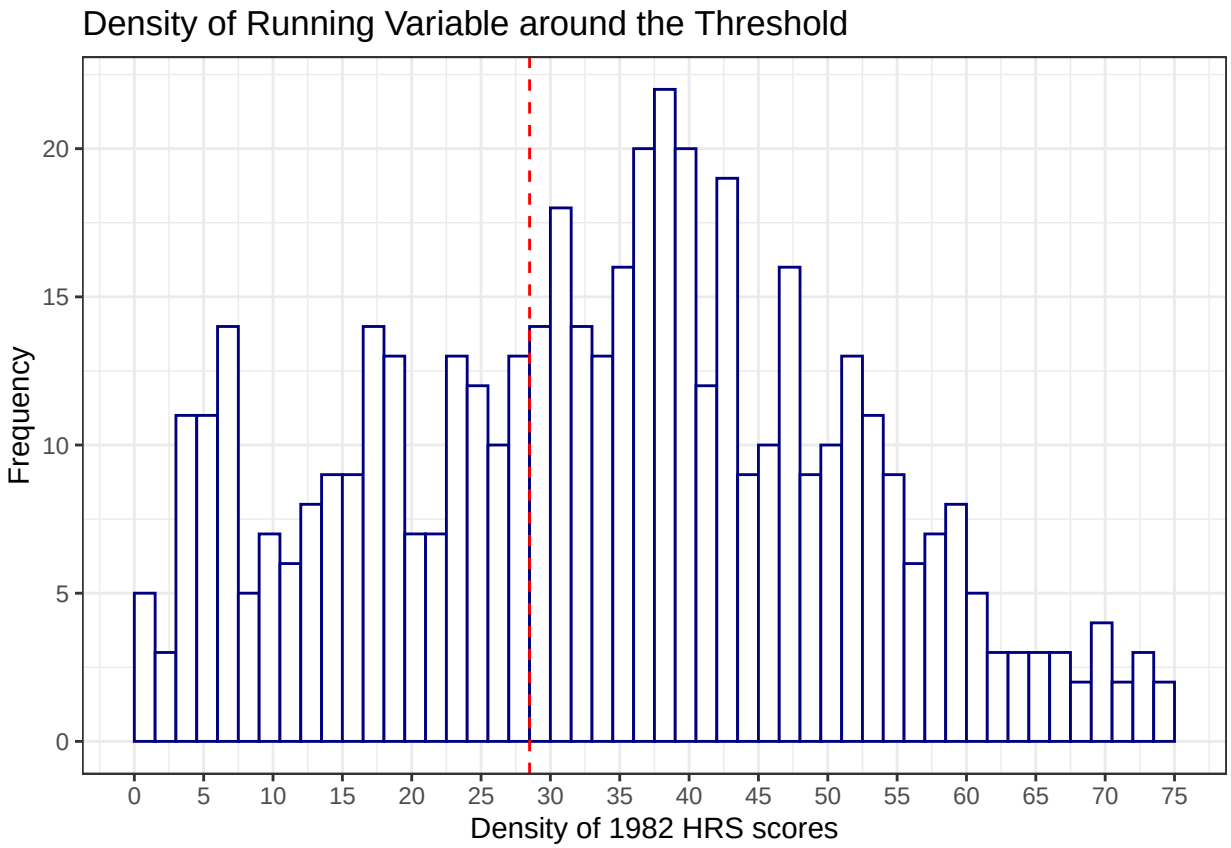
This fact makes it more likely that our assumptions hold. In particular, if none of the individuals involved knew the cutoff threshold score, it is less likely they were able to manipulate the test results to make certain census tracts be above (or below) the threshold. Even if individuals had an incentive 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?

```
## histogram of the density of 1982 HRS scores
ggplot(data, aes(x = hrs_82)) +
  geom_histogram(binwidth = 1.5, boundary = 0, closed = "left", col = "navy", fill = "white") +
  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("Frequency") +
  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 average of the treatment variable (NPL assignment) for each bin
npl2000_bin =tapply(data$npl2000, data$hrs_82_bin, mean)

# generate average outcome in each bin
lnmdvalhs0_nbr_bin = tapply(data$lnmdvalhs0_nbr, data$hrs_82_bin, mean)

# regression fitted on data below the cutoff
below_lm <- lm(lnmdvalhs0_nbr_bin ~ bins.midpoint, data.frame(lnmdvalhs0_nbr_bin, bins.midpoint)[1:19,])
summary(below_lm)
```

```
##
## Call:
## lm(formula = lnmdvalhs0_nbr_bin ~ bins.midpoint, data = data.frame(lnmdvalhs0_nbr_bin,
```

```
##     bins.midpoint)[1:19, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32629 -0.08308  0.03012  0.08214  0.30266
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.486342   0.080292 143.057   <2e-16 ***
## bins.midpoint  0.007585   0.004881   1.554    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1748 on 17 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.07284
## F-statistic: 2.414 on 1 and 17 DF,  p-value: 0.1387
```

```r
# regression fitted on data above the cutoff
above_lm <- lm(lnmdvalhs0_nbr_bin ~ bins.midpoint, data.frame(lnmdvalhs0_nbr_bin, bins.midpoint)[20:50,])
summary(above_lm)
```

```
##
## Call:
## lm(formula = lnmdvalhs0_nbr_bin ~ bins.midpoint, data = data.frame(lnmdvalhs0_nbr_bin,
##     bins.midpoint)[20:50, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44550 -0.10299 -0.02503  0.11681  0.31197
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.657923   0.124356  93.746   <2e-16 ***
## bins.midpoint  0.001049   0.002326   0.451    0.655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1738 on 29 degrees of freedom
## Multiple R-squared:  0.006959,   Adjusted R-squared:  -0.02728
## F-statistic: 0.2032 on 1 and 29 DF,  p-value: 0.6555
```

We estimate the treatment effect $\hat{\tau} = \hat{\alpha}_r - \hat{\alpha}_l = 11.658 - 11.486 = 0.172$, the difference in the estimated intercepts from our local linear regressions above and below the threshold. This is suggestive evidence that being above the threshold, and therefore being more likely to be placed on the NPL, is associated with higher mean housing prices in 2000. Of course, a drawback of fitting separate local linear regressions on either side of the threshold is that we cannot conduct statistical inference on our estimated treatment effect.

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 other covariates. Now we estimate this first stage regression, including controls for population density, education (college educated), children, poverty rate, and home characteristics (no full kitchen, 3 or more bedrooms, mobile).

```
first_stage <- lm(npl2000 ~ above_28pt5 + pop_den8_nbr + ba_or_better8_nbr + child8_nbr + povrat8_nbr +
                  nofullkitchen80_nbr + bedrms3_80occ_nbr + mobile80occ_nbr,
                data = data)
summary(first_stage)
```

```
##
## Call:
## lm(formula = npl2000 ~ above_28pt5 + pop_den8_nbr + ba_or_better8_nbr +
##     child8_nbr + povrat8_nbr + nofullkitchen80_nbr + bedrms3_80occ_nbr +
##     mobile80occ_nbr, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99305 -0.13976 -0.00303  0.01608  0.89830
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.613e-01  1.125e-01   2.323   0.0206 *
## above_28pt5          8.116e-01  2.311e-02  35.120   <2e-16 ***
## pop_den8_nbr        -4.754e-06  2.985e-06  -1.592   0.1120
## ba_or_better8_nbr    1.370e-01  1.721e-01   0.796   0.4264
## child8_nbr          -1.839e-01  2.699e-01  -0.681   0.4960
## povrat8_nbr         -1.979e-02  2.250e-01  -0.088   0.9300
## nofullkitchen80_nbr -8.361e-01  6.089e-01  -1.373   0.1704
## bedrms3_80occ_nbr   -4.647e-02  1.462e-01  -0.318   0.7507
## mobile80occ_nbr      1.730e-02  1.547e-01   0.112   0.9110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2372 on 474 degrees of freedom
## Multiple R-squared:  0.7431, Adjusted R-squared:  0.7388
## F-statistic: 171.4 on 8 and 474 DF,  p-value: < 2.2e-16
```

As expected, having HRS above 28.5 is strongly predictive of NPL status. Now we rerun our IV analysis but focusing on census tracts with HRS between 16.5 and 40.5

```
data_narrow <- data[data$hrs_82 >= 16.5 & data$hrs_82 <= 40.5,]
first_stage <- lm(npl2000 ~ above_28pt5 + pop_den8_nbr + ba_or_better8_nbr + child8_nbr + povrat8_nbr +
                  nofullkitchen80_nbr + bedrms3_80occ_nbr + mobile80occ_nbr,
                data = data_narrow)
summary(first_stage)
```

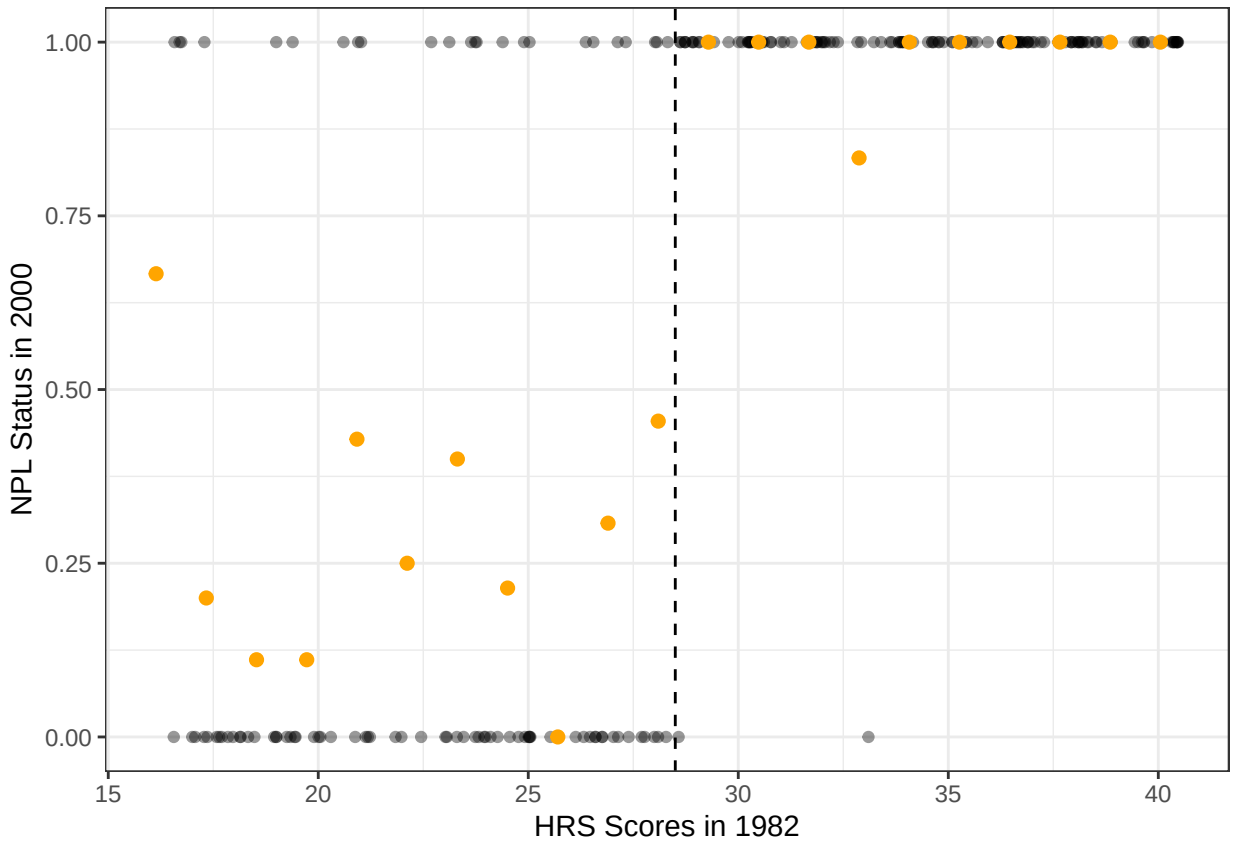
```
##
## Call:
## lm(formula = npl2000 ~ above_28pt5 + pop_den8_nbr + ba_or_better8_nbr +
##     child8_nbr + povrat8_nbr + nofullkitchen80_nbr + bedrms3_80occ_nbr +
##     mobile80occ_nbr, data = data_narrow)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98494 -0.22997 -0.00238  0.02202  0.86897
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.898e-01  2.017e-01   1.436   0.1523
## above_28pt5          7.080e-01  4.070e-02  17.396   <2e-16 ***
## pop_den8_nbr        -6.086e-06  5.464e-06  -1.114   0.2666
```

```
## ba_or_better8_nbr    1.010e-01  2.941e-01   0.344   0.7315
## child8_nbr          -3.404e-01  4.785e-01  -0.711   0.4776
## povrat8_nbr          3.785e-01  3.884e-01   0.975   0.3309
## nofullkitchen80_nbr -1.768e+00  9.395e-01  -1.882   0.0612 .
## bedrms3_80occ_nbr    1.770e-01  2.701e-01   0.655   0.5129
## mobile80occ_nbr     -9.817e-02  3.098e-01  -0.317   0.7516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.296 on 217 degrees of freedom
## Multiple R-squared:  0.5968, Adjusted R-squared:  0.5819
## F-statistic: 40.14 on 8 and 217 DF,  p-value: < 2.2e-16
```

Here again the threshold is strongly predictve of NPL status.

(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.

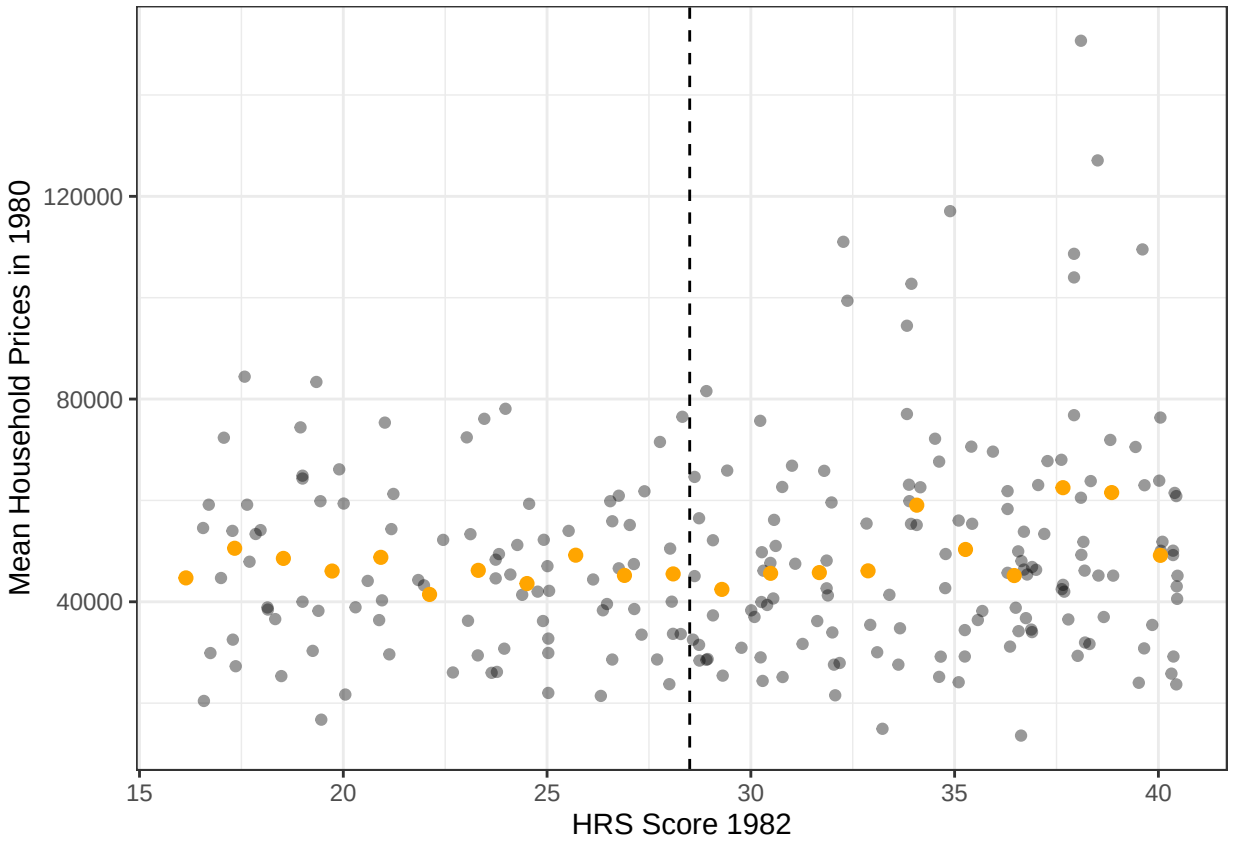
```
(ggplot(data_narrow, aes(x=data_narrow$hrs_82,y=data_narrow$npl2000)) +
  geom_point(alpha = 0.4) +
  stat_summary_bin(fun='mean', bins=20,
                   color='orange', size=2, geom='point')) +
  geom_vline(xintercept=28.5, linetype="dashed") +
  ylab("NPL Status in 2000") +
  xlab("HRS Scores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

```
(ggplot(data_narrow, aes(x=data_narrow$hrs_82,y=data_narrow$meanhs8)) +
  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. If anything, a higher HRS appears to be correlated with slightly higher housing values, which would cause our estimates to be downward biased, if anything. All in all, the values are largely comparable across this range.

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 mean 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 sites between 16.5 and 40.5.

```
iv_reg_full <- iv_robust(lnmdvalhs0_nbr~ npl2000 + pop_den8_nbr + ba_or_better8_nbr + child8_nbr + povrat8_nb
                nofullkitchen80_nbr + bedrms3_80occ_nbr + mobile80occ_nbr |
                  above_28pt5 + pop_den8_nbr + ba_or_better8_nbr + child8_nbr + povrat8_nbr +
                nofullkitchen80_nbr + bedrms3_80occ_nbr + mobile80occ_nbr,
                data = data, se_type = "HC1")
summary(iv_reg_full, digits=4)
```

```
##
## Call:
## iv_robust(formula = lnmdvalhs0_nbr ~ npl2000 + pop_den8_nbr +
##     ba_or_better8_nbr + child8_nbr + povrat8_nbr + nofullkitchen80_nbr +
##     bedrms3_80occ_nbr + mobile80occ_nbr | above_28pt5 + pop_den8_nbr +
##     ba_or_better8_nbr + child8_nbr + povrat8_nbr + nofullkitchen80_nbr +
##     bedrms3_80occ_nbr + mobile80occ_nbr, data = data, se_type = "HC1")
##
## Standard error type:  HC1
##
## Coefficients:
##                       Estimate Std. Error  t value   Pr(>|t|)   CI Lower
## (Intercept)          1.143e+01  1.974e-01  57.9061 4.087e-217  1.104e+01
## npl2000              2.677e-02  3.653e-02   0.7327  4.641e-01 -4.502e-02
## pop_den8_nbr         2.461e-05  3.799e-06   6.4779  2.332e-10  1.714e-05
## ba_or_better8_nbr    3.083e+00  2.583e-01  11.9351  6.809e-29  2.576e+00
## child8_nbr           1.727e+00  4.074e-01   4.2382  2.709e-05  9.262e-01
## povrat8_nbr         -3.028e+00  2.845e-01 -10.6435  7.241e-24 -3.586e+00
## nofullkitchen80_nbr  1.666e+00  7.104e-01   2.3455  1.941e-02  2.703e-01
## bedrms3_80occ_nbr   -1.045e+00  2.084e-01  -5.0119  7.624e-07 -1.454e+00
## mobile80occ_nbr      5.928e-01  2.187e-01   2.7107  6.958e-03  1.631e-01
##                       CI Upper  DF
## (Intercept)          1.182e+01 474
## npl2000              9.856e-02 474
## pop_den8_nbr         3.207e-05 474
## ba_or_better8_nbr    3.591e+00 474
## child8_nbr           2.527e+00 474
## povrat8_nbr         -2.469e+00 474
## nofullkitchen80_nbr  3.062e+00 474
## bedrms3_80occ_nbr   -6.351e-01 474
## mobile80occ_nbr      1.023e+00 474
##
## Multiple R-squared:  0.588 , Adjusted R-squared:  0.581
## F-statistic: 86.83 on 8 and 474 DF,  p-value: < 2.2e-16
```

```
iv_reg_narrow <- iv_robust(lnmdvalhs0_nbr ~ npl2000 + pop_den8_nbr + ba_or_better8_nbr + child8_nbr + povrat8
                nofullkitchen80_nbr + bedrms3_80occ_nbr + mobile80occ_nbr |
                  above_28pt5 + pop_den8_nbr + ba_or_better8_nbr + child8_nbr + povrat8_nbr +
                nofullkitchen80_nbr + bedrms3_80occ_nbr + mobile80occ_nbr,
                data = data_narrow, se_type = "HC1")
summary(iv_reg_narrow, digits = 4)
```

```
##
## Call:
## iv_robust(formula = lnmdvalhs0_nbr ~ npl2000 + pop_den8_nbr +
##     ba_or_better8_nbr + child8_nbr + povrat8_nbr + nofullkitchen80_nbr +
##     bedrms3_80occ_nbr + mobile80occ_nbr | above_28pt5 + pop_den8_nbr +
##     ba_or_better8_nbr + child8_nbr + povrat8_nbr + nofullkitchen80_nbr +
##     bedrms3_80occ_nbr + mobile80occ_nbr, data = data_narrow,
##
## Standard error type:  HC1
##
## Coefficients:
##                       Estimate Std. Error t value  Pr(>|t|)   CI Lower
## (Intercept)          1.150e+01  3.578e-01 32.1448 1.797e-84  1.080e+01
## npl2000             -3.255e-02  5.760e-02 -0.5651 5.726e-01 -1.461e-01
## pop_den8_nbr         2.098e-05  4.423e-06  4.7425 3.824e-06  1.226e-05
## ba_or_better8_nbr    2.613e+00  3.336e-01  7.8318 2.113e-13  1.955e+00
## child8_nbr           1.719e+00  6.069e-01  2.8320 5.062e-03  5.225e-01
## povrat8_nbr         -3.340e+00  3.384e-01 -9.8696 3.244e-19 -4.007e+00
```

```
## nofullkitchen80_nbr  2.807e+00  7.378e-01  3.8048 1.846e-04  1.353e+00
## bedrms3_80occ_nbr   -9.827e-01  3.578e-01 -2.7469 6.522e-03 -1.688e+00
## mobile80occ_nbr      5.265e-01  3.666e-01  1.4361 1.524e-01 -1.961e-01
##                       CI Upper  DF
## (Intercept)          1.221e+01 217
## npl2000              8.098e-02 217
## pop_den8_nbr         2.969e-05 217
## ba_or_better8_nbr    3.270e+00 217
## child8_nbr           2.915e+00 217
## povrat8_nbr         -2.673e+00 217
## nofullkitchen80_nbr  4.262e+00 217
## bedrms3_80occ_nbr   -2.776e-01 217
## mobile80occ_nbr      1.249e+00 217
##
## Multiple R-squared:  0.5548 ,    Adjusted R-squared:  0.5383
## F-statistic:  44.3 on 8 and 217 DF,  p-value: < 2.2e-16
```

Question 5: Conclusion