

# ARE213 Problem Set #3

Peter Alstone & Frank Proulx

December 10, 2013

## Part A: Linear models to motivate RD

### (i) LM results comparison

Using a series of linear models (with heteroskedasticity consistent “robust” standard errors), we find that for a range of model formulations there is a significant effect on housing price from the presence of hazardous waste cleanup sites with increased housing values in places with cleanup. The coefficient for the hazardous waste cleanup indicator variable (npl2000) takes a wide range of values depending on which additional explanatory variables are included in the model, from 0.04 (i.e., approximately a 4% increase) for the simple model only including 1980 housing values and npl2000 to estimate 2000 housing values, to 0.09 for a model including both housing and demographic characteristics.

**Requirements for Unbiased Estimates:** For our estimates to be unbiased we would need to include all of the potential sources of variation in housing price in a linear model. A particular challenge is that there are very few sites with NPL2000 status (only 2% of sites), so while the overall sample size is large there is very little support for estimates related to NPL2000 status compared to other covariates. The overlap assumption must hold for the regression to be successful.

### (ii) Comparing covariates

We compare the covariates between census tracts and sites in a series of contingency tables and find that there are wide disparities between census tracts with and without NPL2000 status. This erodes confidence that there is support in the data to use tract-level linear regression models, since the overlap assumption may be violated from wide differences in the other characteristics on the tract level. On the site level, simply comparing over / under the trigger limit for the national priorities list (HRS score of 28.5) seems to solve some but not all the problems with overlap. While many covariates cannot be said to come from different distributions there are still some that have significant differences. Narrowing into a window from 16.5 - 40.5 (with the 28.5 dividing line) results in comparisons for which the hypothesis that the covariates are from the same distribution is not rejected. Overall, these comparisons motivate the regression discontinuity design. By narrowing in on a region where overlap in distribution for the covariates holds we have a fighting chance to identify a treatment effect, albeit with difficulty in establishing external validity.

## Part B: RDD setup

### (i) HRS as running variable?

To be a running variable HRS need to be continuous and not subject to manipulation around the boundary value. If the variable is as good as random around the cutoff (which is based solely on its value) we can use

Table 1: Linear models for effect of NPL(2000) on housing value (with many additional state fixed effects omitted)

	simple model	+housing char.	+demographics	+state fixed effects
	(1)	(2)	(3)	(4)
npl2000	0.040*** (0.012)	0.055*** (0.012)	0.090*** (0.010)	0.068*** (0.009)
lnmeanhs8	0.856*** (0.011)	0.866*** (0.018)	0.619*** (0.022)	0.514*** (0.022)
firestoveheat80		0.074*** (0.020)	0.182*** (0.023)	0.230*** (0.033)
nofullkitchen80		-1.776*** (0.176)	-0.751*** (0.164)	-0.559*** (0.152)
zerofullbath80		1.243*** (0.139)	1.044*** (0.124)	0.863*** (0.116)
bedrms1_80occ		0.421* (0.249)	0.404* (0.237)	0.240 (0.234)
bedrms2_80occ		-0.436* (0.229)	0.156 (0.216)	-0.004 (0.214)
bedrms3_80occ		-0.524** (0.230)	-0.147 (0.217)	-0.153 (0.214)
bedrms4_80occ		-0.111 (0.226)	0.004 (0.217)	-0.213 (0.214)
bedrms5_80occ		0.721*** (0.231)	0.732*** (0.222)	0.430* (0.220)
blt0_1yrs80occ		-0.216*** (0.045)	-0.010 (0.044)	0.109** (0.045)
blt2_5yrs80occ		-0.295*** (0.029)	0.011 (0.028)	0.039 (0.026)
blt6_10yrs80occ		-0.271*** (0.021)	-0.048** (0.021)	0.002 (0.021)
blt10_20yrs80occ		-0.242*** (0.017)	-0.136*** (0.015)	-0.123*** (0.014)
blt20_30yrs80occ		-0.191*** (0.017)	-0.181*** (0.014)	-0.156*** (0.013)
blt30_40yrs80occ		-0.190*** (0.026)	-0.121*** (0.025)	-0.104*** (0.023)
occupied80		0.730*** (0.050)	0.242*** (0.046)	-0.093** (0.044)
pop_den8			0.00001*** (0.00000)	0.00001*** (0.00000)
shrbk8			-0.161*** (0.014)	-0.058*** (0.013)
shrhsp8			-0.329*** (0.021)	-0.100*** (0.022)
child8			-0.630*** (0.058)	-0.431*** (0.052)
old8			-0.737*** (0.047)	-0.447*** (0.044)
shrfors8			1.377*** (0.048)	0.567*** (0.041)
ffh8			-0.006 (0.034)	-0.084*** (0.032)
smhse8			0.407*** (0.022)	0.323*** (0.022)
hsdrop8			0.010 (0.025)	0.042* (0.024)
no_hs_diploma8			-0.537*** (0.039)	-0.262*** (0.034)
ba_or_better8			0.112*** (0.034)	0.450*** (0.035)
unemp8			-0.654*** (0.071)	-1.420*** (0.076)
povrat8			-0.275*** (0.051)	0.118** (0.048)
welfare8			1.271*** (0.070)	0.284*** (0.067)
avhhin8			0.00001*** (0.00000)	0.00001*** (0.00000)
as.factor(statefips)2				-0.129*** (0.027)
as.factor(statefips)4				0.011 (0.015)
as.factor(statefips)5				-0.150*** (0.025)
as.factor(statefips)6				0.340*** (0.017)
as.factor(statefips)8				0.207*** (0.015)
as.factor(statefips)9				0.157*** (0.015)
as.factor(statefips)10				0.230*** (0.018)
as.factor(statefips)11				0.102*** (0.024)
as.factor(statefips)12				-0.005 (0.013)
as.factor(statefips)13				0.182*** (0.015)
as.factor(statefips)15				0.081** (0.038)
as.factor(statefips)16				0.039* (0.020)

Notes:

\*\*\*Significant at the 1 percent level.

\*\*Significant at the 5 percent level.

\*Significant at the 10 percent level.

Table 2: Contingency table for a range of factors by npl2000 status

	N	0 N = 47260	1 N = 985	Combined N = 48245	Test Statistic
npl1990 : 0	48245	100% (47260)	24% ( 239)	98% (47499)	$\chi^2_1 = 36355, P < 0.001^1$
1		0% ( 0)	76% ( 746)	2% ( 746)	
pop_den8	48245	580 2677 6178	147 522 1701	548 2605 6080	$F_{1,48243} = 525, P < 0.001^2$
shrblk8	48245	0.00 0.02 0.08	0.00 0.02 0.07	0.00 0.02 0.08	$F_{1,48243} = 0.35, P = 0.55^2$
shrhsp8	48245	0.01 0.02 0.06	0.01 0.01 0.03	0.01 0.02 0.06	$F_{1,48243} = 39, P < 0.001^2$
child8	48245	0.24 0.29 0.33	0.26 0.30 0.33	0.24 0.29 0.33	$F_{1,48243} = 35, P < 0.001^2$
shrfor8	48245	0.02 0.04 0.08	0.02 0.03 0.06	0.02 0.04 0.08	$F_{1,48243} = 44, P < 0.001^2$
ffh8	48245	0.10 0.15 0.24	0.09 0.13 0.20	0.10 0.15 0.24	$F_{1,48243} = 40, P < 0.001^2$
smhse8	48245	0.42 0.53 0.63	0.47 0.56 0.64	0.42 0.53 0.63	$F_{1,48243} = 34, P < 0.001^2$
hsdrop8	48245	0.05 0.11 0.19	0.06 0.12 0.19	0.05 0.11 0.19	$F_{1,48243} = 3.8, P = 0.051^2$
no_hs_diploma8	48245	0.18 0.29 0.43	0.23 0.33 0.43	0.19 0.29 0.43	$F_{1,48243} = 42, P < 0.001^2$
ba_or_better8	48245	0.08 0.14 0.24	0.07 0.11 0.18	0.08 0.14 0.24	$F_{1,48243} = 52, P < 0.001^2$
unemprrt8	48245	0.04 0.06 0.08	0.04 0.06 0.08	0.04 0.06 0.08	$F_{1,48243} = 22, P < 0.001^2$
povrat8	48245	0.05 0.08 0.14	0.05 0.08 0.13	0.05 0.08 0.14	$F_{1,48243} = 0.02, P = 0.9^2$
welfare8	48245	0.03 0.05 0.09	0.03 0.05 0.09	0.03 0.05 0.09	$F_{1,48243} = 4.2, P = 0.041^2$
favinc8	48245	18786 22882 27697	18943 22085 25676	18789 22863 27660	$F_{1,48243} = 15, P < 0.001^2$
avhhin8	48245	16349 20383 25211	16766 19957 23589	16358 20371 25166	$F_{1,48243} = 5.7, P = 0.017^2$
meanrnt80	48115	223 268 324	217 256 303	223 268 323	$F_{1,48113} = 27, P < 0.001^2$
mdvalls9	48245	43929 69394 125500	48500 72700 130600	44000 69400 125600	$F_{1,48243} = 6.1, P = 0.014^2$
meanrnt9	48190	390 491 636	378 471 620	389 491 635	$F_{1,48188} = 6.3, P = 0.012^2$
mdvalls0	48245	82500 120700 178000	85400 120400 166200	82600 120700 177800	$F_{1,48243} = 0.28, P = 0.6^2$
meanrnt0	48127	520 646 822	515 621 800	520 645 821	$F_{1,48125} = 5, P = 0.025^2$
tothsun8	48245	874 1278 1735	937 1343 1807	875 1280 1737	$F_{1,48243} = 9, P = 0.003^2$
ownocc8	48245	448 748 1089	566 878 1212	450 751 1091	$F_{1,48243} = 60, P < 0.001^2$
owner_occupied80	48245	0.48 0.67 0.79	0.58 0.71 0.80	0.49 0.67 0.79	$F_{1,48243} = 38, P < 0.001^2$
bltlast5yrs80	48245	0.02 0.09 0.22	0.05 0.12 0.20	0.02 0.09 0.22	$F_{1,48243} = 24, P < 0.001^2$
bltlast10yrs80	48245	0.07 0.22 0.45	0.13 0.26 0.40	0.07 0.22 0.45	$F_{1,48243} = 14, P < 0.001^2$
firestoveheat80	48245	0.00 0.01 0.04	0.01 0.03 0.07	0.00 0.01 0.05	$F_{1,48243} = 132, P < 0.001^2$
noaircond80	48245	0.17 0.40 0.66	0.30 0.47 0.68	0.17 0.40 0.66	$F_{1,48243} = 58, P < 0.001^2$
nofullkitchen80	48245	0.00 0.01 0.02	0.01 0.01 0.02	0.00 0.01 0.02	$F_{1,48243} = 20, P < 0.001^2$
zerofullbath80	48245	0.00 0.01 0.03	0.01 0.02 0.03	0.00 0.01 0.03	$F_{1,48243} = 48, P < 0.001^2$
northeast : 0	48245	78% (36651)	62% ( 611)	77% (37262)	$\chi^2_1 = 132, P < 0.001^1$
1		22% (10609)	38% ( 374)	23% (10983)	
midwest : 0	48245	77% (36338)	78% ( 770)	77% (37108)	$\chi^2_1 = 0.89, P = 0.34^1$
1		23% (10922)	22% ( 215)	23% (11137)	
south : 0	48245	69% (32425)	76% ( 753)	69% (33178)	$\chi^2_1 = 28, P < 0.001^1$
1		31% (14835)	24% ( 232)	31% (15067)	
west : 0	48245	77% (36366)	83% ( 821)	77% (37187)	$\chi^2_1 = 22, P < 0.001^1$
1		23% (10894)	17% ( 164)	23% (11058)	
meanhs8	48245	38270 52659 73074	38213 49126 64321	38269 52576 72906	$F_{1,48243} = 20, P < 0.001^2$
bedrms02.80	48245	0.32 0.46 0.62	0.33 0.43 0.55	0.32 0.45 0.62	$F_{1,48243} = 10, P = 0.001^2$
bedrms34.80	48245	0.36 0.52 0.65	0.43 0.55 0.63	0.36 0.52 0.65	$F_{1,48243} = 9.7, P = 0.002^2$
detach80	43074	0.45 0.70 0.84	0.57 0.73 0.83	0.46 0.70 0.84	$F_{1,43072} = 16, P < 0.001^2$
bedrms0.80occ	48245	0 0 0	0 0 0	0 0 0	$F_{1,48243} = 0.22, P = 0.64^2$
bedrms1.80occ	48245	0.01 0.03 0.06	0.02 0.03 0.06	0.01 0.03 0.06	$F_{1,48243} = 7.4, P = 0.007^2$
bedrms2.80occ	48245	0.15 0.25 0.36	0.19 0.27 0.35	0.16 0.25 0.36	$F_{1,48243} = 16, P < 0.001^2$
bedrms3.80occ	48245	0.39 0.48 0.57	0.43 0.49 0.55	0.39 0.48 0.57	$F_{1,48243} = 1.2, P = 0.27^2$
bedrms4.80occ	48245	0.09 0.14 0.22	0.10 0.15 0.21	0.09 0.14 0.22	$F_{1,48243} = 0.27, P = 0.61^2$
bedrms5.80occ	48245	0.01 0.02 0.05	0.01 0.03 0.04	0.01 0.02 0.05	$F_{1,48243} = 1.5, P = 0.22^2$
blt0.1yrs80occ	48245	0.00 0.01 0.05	0.01 0.02 0.05	0.00 0.01 0.05	$F_{1,48243} = 44, P < 0.001^2$
blt2.5yrs80occ	48245	0.01 0.06 0.17	0.02 0.09 0.16	0.01 0.06 0.17	$F_{1,48243} = 42, P < 0.001^2$
blt6.10yrs80occ	48245	0.01 0.08 0.19	0.05 0.12 0.19	0.01 0.09 0.19	$F_{1,48243} = 52, P < 0.001^2$
blt10.20yrs80occ	48245	0.07 0.16 0.26	0.12 0.18 0.25	0.07 0.16 0.26	$F_{1,48243} = 29, P < 0.001^2$
blt20.30yrs80occ	48245	0.06 0.14 0.26	0.10 0.16 0.24	0.07 0.14 0.26	$F_{1,48243} = 31, P < 0.001^2$
blt30.40yrs80occ	48245	0.03 0.07 0.14	0.05 0.08 0.13	0.03 0.07 0.14	$F_{1,48243} = 21, P < 0.001^2$
blt40.yrs80occ	48245	0.03 0.14 0.43	0.07 0.18 0.32	0.03 0.14 0.42	$F_{1,48243} = 11, P = 0.001^2$
detach80occ	48245	0.86 0.96 0.99	0.83 0.93 0.98	0.86 0.96 0.99	$F_{1,48243} = 40, P < 0.001^2$
attach80occ	48245	0.00 0.01 0.04	0.00 0.01 0.02	0.00 0.01 0.04	$F_{1,48243} = 40, P < 0.001^2$
mobile80occ	48245	0.00 0.00 0.06	0.00 0.04 0.13	0.00 0.00 0.06	$F_{1,48243} = 238, P < 0.001^2$
occupied80	48245	0.92 0.95 0.97	0.93 0.95 0.97	0.92 0.95 0.97	$F_{1,48243} = 3.4, P = 0.066^2$
bltmore30.80	48245	0.08 0.27 0.56	0.17 0.31 0.46	0.08 0.28 0.55	$F_{1,48243} = 8.8, P = 0.003^2$
nbr_dummy : 0	48245	89% (41989)	56% ( 551)	88% (42540)	$\chi^2_1 = 1002, P < 0.001^1$
1		11% ( 5271)	44% ( 434)	12% ( 5705)	

$a\ b\ c$  represent the lower quartile  $a$ , the median  $b$ , and the upper quartile  $c$  for continuous variables.

$N$  is the number of non-missing values.

Numbers after percents are frequencies.

Tests used:

<sup>1</sup>Pearson test; <sup>2</sup>Wilcoxon test

Table 3: Contingency table by HRS test status (over/under 28.5)

	N	FALSE <i>N</i> = 181	TRUE <i>N</i> = 306	Combined <i>N</i> = 487	Test Statistic
hrs.82	487	7.5 16.5 23.1	36.3 42.3 51.9	20.2 34.7 46.0	$F_{1,485} = 1135, P < 0.001^1$
npl1990 : 0	487	87% (158)	1% ( 3)	33% (161)	$\chi_1^2 = 383, P < 0.001^2$
1		13% ( 23)	99% (303)	67% (326)	
npl2000 : 0	487	84% (152)	1% ( 3)	32% (155)	$\chi_1^2 = 361, P < 0.001^2$
1		16% ( 29)	99% (303)	68% (332)	
pop_den8	487	146 533 1939	146 483 1415	145 504 1568	$F_{1,485} = 0.4, P = 0.53^1$
shrbk8	487	0.00 0.01 0.06	0.00 0.02 0.05	0.00 0.01 0.05	$F_{1,485} = 0.45, P = 0.5^1$
shrhsp8	487	0.00 0.01 0.03	0.00 0.01 0.03	0.00 0.01 0.03	$F_{1,485} = 0.27, P = 0.6^1$
child8	487	0.26 0.30 0.33	0.27 0.30 0.33	0.26 0.30 0.33	$F_{1,485} = 0.01, P = 0.93^1$
shrfor8	487	0.01 0.03 0.05	0.02 0.03 0.06	0.01 0.03 0.06	$F_{1,485} = 4.7, P = 0.03^1$
ffh8	487	0.11 0.15 0.21	0.08 0.13 0.19	0.09 0.14 0.20	$F_{1,485} = 7.4, P = 0.007^1$
smhse8	487	0.52 0.61 0.68	0.48 0.57 0.66	0.50 0.59 0.66	$F_{1,485} = 8.5, P = 0.004^1$
hsdrop8	487	0.07 0.13 0.20	0.06 0.11 0.18	0.07 0.12 0.19	$F_{1,485} = 2.9, P = 0.09^1$
no_hs_diploma8	487	0.30 0.39 0.50	0.24 0.34 0.42	0.26 0.36 0.46	$F_{1,485} = 20, P < 0.001^1$
ba_or_better8	487	0.05 0.08 0.13	0.07 0.11 0.18	0.06 0.10 0.16	$F_{1,485} = 22, P < 0.001^1$
unemprt8	487	0.05 0.07 0.10	0.04 0.06 0.09	0.05 0.07 0.09	$F_{1,485} = 11, P < 0.001^1$
povrat8	487	0.06 0.09 0.14	0.05 0.07 0.13	0.05 0.08 0.13	$F_{1,485} = 3.9, P = 0.048^1$
welfare8	487	0.04 0.07 0.09	0.04 0.05 0.09	0.04 0.06 0.09	$F_{1,485} = 9.7, P = 0.002^1$
favinc8	487	18744 21026 24470	19054 22301 26440	18906 21693 25444	$F_{1,485} = 4.8, P = 0.029^1$
avhhi8	487	16862 19578 22230	17198 20209 23805	16939 19869 23172	$F_{1,485} = 3.9, P = 0.05^1$
meanrnt80	484	208 242 284	218 256 309	214 249 300	$F_{1,482} = 9.3, P = 0.002^1$
mdvalhs9	487	43800 58300 116100	50335 73500 131175	46800 67407 125450	$F_{1,485} = 8.4, P = 0.004^1$
meanrnt9	487	352 412 569	380 470 629	369 449 596	$F_{1,485} = 13, P < 0.001^1$
mdvalhs0	487	73900 101000 145400	87850 121200 161800	82150 114400 156650	$F_{1,485} = 14, P < 0.001^1$
meanrnt0	485	470 560 700	520 618 824	501 594 777	$F_{1,483} = 18, P < 0.001^1$
tothsun8	487	891 1273 1677	905 1304 1753	902 1292 1728	$F_{1,485} = 0.24, P = 0.63^1$
ownocc8	487	571 832 1157	585 872 1210	576 860 1180	$F_{1,485} = 0.12, P = 0.73^1$
owner_occupied80	487	0.61 0.71 0.79	0.61 0.72 0.80	0.61 0.72 0.80	$F_{1,485} = 0.32, P = 0.57^1$
bltlast5yrs80	487	0.02 0.10 0.17	0.05 0.12 0.20	0.04 0.11 0.19	$F_{1,485} = 5.8, P = 0.017^1$
bltlast10yrs80	487	0.09 0.22 0.34	0.14 0.25 0.41	0.12 0.24 0.38	$F_{1,485} = 7.5, P = 0.006^1$
firestoveheat80	487	0.01 0.02 0.07	0.01 0.02 0.06	0.01 0.02 0.06	$F_{1,485} = 0.05, P = 0.82^1$
noaircond80	487	0.34 0.50 0.70	0.29 0.47 0.67	0.31 0.48 0.68	$F_{1,485} = 1.4, P = 0.23^1$
nofullkitchen80	487	0.01 0.01 0.03	0.00 0.01 0.02	0.00 0.01 0.02	$F_{1,485} = 2.6, P = 0.1^1$
zerofullbath80	487	0.01 0.02 0.04	0.01 0.02 0.03	0.01 0.02 0.03	$F_{1,485} = 5.3, P = 0.021^1$
northeast : 0	487	67% (121)	52% (160)	58% (281)	$\chi_1^2 = 9.9, P = 0.002^2$
1		33% ( 60)	48% (146)	42% (206)	
midwest : 0	487	65% (118)	77% (237)	73% (355)	$\chi_1^2 = 8.7, P = 0.003^2$
1		35% ( 63)	23% ( 69)	27% (132)	
south : 0	487	78% (142)	81% (247)	80% (389)	$\chi_1^2 = 0.36, P = 0.55^2$
1		22% ( 39)	19% ( 59)	20% ( 98)	
west : 0	487	90% (162)	90% (274)	90% (436)	$\chi_1^2 = 0, P = 0.99^2$
1		10% ( 19)	10% ( 32)	10% ( 51)	
meanhs8	487	30749 41910 55157	37082 48084 62641	35536 46152 59844	$F_{1,485} = 16, P < 0.001^1$
bedrms02.80	487	0.37 0.46 0.56	0.33 0.43 0.54	0.35 0.44 0.54	$F_{1,485} = 4.7, P = 0.031^1$
bedrms34.80	487	0.42 0.52 0.61	0.44 0.55 0.63	0.43 0.54 0.62	$F_{1,485} = 3.7, P = 0.055^1$
detach80	400	0.55 0.74 0.84	0.58 0.74 0.86	0.57 0.74 0.85	$F_{1,398} = 1.4, P = 0.23^1$
bedrms0.80occ	487	0 0 0	0 0 0	0 0 0	$F_{1,485} = 0.02, P = 0.88^1$
bedrms1.80occ	487	0.02 0.03 0.06	0.02 0.03 0.05	0.02 0.03 0.05	$F_{1,485} = 0.44, P = 0.51^1$
bedrms2.80occ	487	0.22 0.30 0.40	0.19 0.26 0.34	0.20 0.28 0.36	$F_{1,485} = 12, P < 0.001^1$
bedrms3.80occ	487	0.42 0.48 0.54	0.44 0.50 0.56	0.43 0.50 0.55	$F_{1,485} = 1.7, P = 0.2^1$
bedrms4.80occ	487	0.09 0.13 0.17	0.10 0.15 0.21	0.09 0.14 0.20	$F_{1,485} = 13, P < 0.001^1$
bedrms5.80occ	487	0.01 0.02 0.04	0.01 0.03 0.04	0.01 0.02 0.04	$F_{1,485} = 7.2, P = 0.008^1$
blt0.1yrs80occ	487	0.00 0.02 0.04	0.01 0.02 0.05	0.00 0.02 0.04	$F_{1,485} = 4.2, P = 0.04^1$
blt2.5yrs80occ	487	0.01 0.06 0.13	0.02 0.09 0.16	0.02 0.08 0.14	$F_{1,485} = 7.5, P = 0.006^1$
blt6.10yrs80occ	487	0.03 0.09 0.16	0.05 0.12 0.20	0.04 0.11 0.18	$F_{1,485} = 6.6, P = 0.011^1$
blt10.20yrs80occ	487	0.10 0.17 0.23	0.13 0.19 0.26	0.12 0.18 0.25	$F_{1,485} = 7.4, P = 0.007^1$
blt20.30yrs80occ	487	0.09 0.15 0.25	0.11 0.16 0.24	0.10 0.16 0.25	$F_{1,485} = 1.6, P = 0.2^1$
blt30.40yrs80occ	487	0.05 0.09 0.16	0.05 0.08 0.12	0.05 0.08 0.14	$F_{1,485} = 2.1, P = 0.15^1$
blt40.yrs80occ	487	0.11 0.23 0.44	0.08 0.17 0.33	0.09 0.19 0.36	$F_{1,485} = 6.8, P = 0.009^1$
detach80occ	487	0.84 0.91 0.97	0.84 0.94 0.99	0.84 0.93 0.99	$F_{1,485} = 2.5, P = 0.12^1$
attach80occ	487	0.00 0.01 0.02	0.00 0.01 0.01	0.00 0.01 0.02	$F_{1,485} = 0.49, P = 0.49^1$
mobile80occ	487	0.00 0.04 0.12	0.00 0.03 0.12	0.00 0.03 0.12	$F_{1,485} = 0.21, P = 0.65^1$
occupied80	487	0.93 0.95 0.97	0.92 0.95 0.97	0.92 0.95 0.97	$F_{1,485} = 0.03, P = 0.87^1$
bltmore30.80	487	0.24 0.38 0.55	0.16 0.30 0.46	0.19 0.32 0.51	$F_{1,485} = 10, P = 0.001^1$
og82list : 1	487	100% (181)	100% (306)	100% (487)	<sup>2</sup>
0		0% ( 0)	0% ( 0)	0% ( 0)	
nbr_dummy : 0	487	80% (144)	66% (203)	71% (347)	$\chi_1^2 = 9.7, P = 0.002^2$
1		20% ( 37)	34% (103)	29% (140)	

*a b c* represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables.

*N* is the number of non-missing values.

Numbers after percents are frequencies.

Tests used:

Table 4: Contingency table by HRS test status (JUST over/under 28.5)

	N	FALSE <i>N</i> = 90	TRUE <i>N</i> = 137	Combined <i>N</i> = 227	Test Statistic
hrs_82	227	19 23 25	32 35 38	24 30 37	$F_{1,225} = 573, P < 0.001^1$
npl1990 : 0	227	78% ( 70)	1% ( 2)	32% ( 72)	$\chi_1^2 = 146, P < 0.001^2$
1		22% ( 20)	99% (135)	68% (155)	
npl2000 : 0	227	73% ( 66)	1% ( 2)	30% ( 68)	$\chi_1^2 = 134, P < 0.001^2$
1		27% ( 24)	99% (135)	70% (159)	
pop_den8	227	114 357 1340	142 427 1178	118 417 1192	$F_{1,225} = 0.14, P = 0.71^1$
shrbk8	227	0.00 0.01 0.03	0.00 0.02 0.05	0.00 0.01 0.05	$F_{1,225} = 2.5, P = 0.12^1$
shrbsp8	227	0.00 0.01 0.02	0.00 0.01 0.02	0.00 0.01 0.02	$F_{1,225} = 0.03, P = 0.87^1$
child8	227	0.26 0.30 0.33	0.26 0.30 0.34	0.26 0.30 0.33	$F_{1,225} = 0.05, P = 0.82^1$
shrfor8	227	0.01 0.03 0.05	0.01 0.03 0.06	0.01 0.03 0.05	$F_{1,225} = 0.09, P = 0.76^1$
ffh8	227	0.10 0.15 0.20	0.09 0.13 0.21	0.09 0.14 0.20	$F_{1,225} = 0.32, P = 0.57^1$
smhse8	227	0.51 0.59 0.65	0.48 0.57 0.66	0.50 0.59 0.65	$F_{1,225} = 0.75, P = 0.39^1$
hsdrop8	227	0.07 0.13 0.20	0.06 0.12 0.18	0.07 0.12 0.20	$F_{1,225} = 1.2, P = 0.27^1$
no_hs_diploma8	227	0.30 0.38 0.48	0.24 0.35 0.44	0.29 0.36 0.45	$F_{1,225} = 3.1, P = 0.077^1$
ba_or_better8	227	0.05 0.10 0.13	0.06 0.10 0.17	0.06 0.10 0.16	$F_{1,225} = 2.3, P = 0.13^1$
unemp8	227	0.05 0.07 0.09	0.05 0.06 0.10	0.05 0.06 0.09	$F_{1,225} = 0.54, P = 0.47^1$
povrat8	227	0.05 0.09 0.13	0.05 0.09 0.14	0.05 0.09 0.13	$F_{1,225} = 0, P = 0.95^1$
welfare8	227	0.04 0.06 0.09	0.04 0.05 0.09	0.04 0.06 0.09	$F_{1,225} = 2.7, P = 0.1^1$
favinc8	227	18951 21005 24908	18603 21513 26037	18843 21343 25095	$F_{1,225} = 0.19, P = 0.67^1$
avhhin8	227	17176 19521 22221	16237 19523 23189	16768 19523 22558	$F_{1,225} = 0, P = 0.98^1$
meanrnt80	227	219 245 285	215 244 293	216 245 287	$F_{1,225} = 0, P = 0.97^1$
mdvalhs9	227	45556 64100 121550	44041 68177 118334	44800 66600 120453	$F_{1,225} = 0.06, P = 0.81^1$
meanrnt9	227	358 422 579	365 432 563	364 431 568	$F_{1,225} = 0.03, P = 0.86^1$
mdvalhs0	227	76600 108750 143600	78300 114400 151800	76600 111700 150200	$F_{1,225} = 0.26, P = 0.61^1$
meanrnt0	225	490 577 701	487 550 710	487 568 704	$F_{1,223} = 0.08, P = 0.78^1$
tothsun8	227	901 1280 1693	886 1308 1713	888 1290 1708	$F_{1,225} = 0.01, P = 0.93^1$
ownocc8	227	594 912 1160	558 879 1238	576 900 1198	$F_{1,225} = 0.5, P = 0.48^1$
owner_occupied80	227	0.62 0.73 0.79	0.61 0.73 0.81	0.61 0.73 0.80	$F_{1,225} = 0, P = 0.95^1$
bltlast5yrs80	227	0.05 0.12 0.20	0.05 0.11 0.19	0.05 0.12 0.20	$F_{1,225} = 0.14, P = 0.71^1$
bltlast10yrs80	227	0.13 0.26 0.36	0.14 0.25 0.38	0.13 0.25 0.37	$F_{1,225} = 0.01, P = 0.92^1$
firestoveheat80	227	0.01 0.04 0.09	0.01 0.02 0.06	0.01 0.03 0.07	$F_{1,225} = 0.67, P = 0.41^1$
noaircond80	227	0.34 0.52 0.71	0.34 0.52 0.68	0.34 0.52 0.70	$F_{1,225} = 0.02, P = 0.88^1$
nofullkitchen80	227	0.01 0.01 0.03	0.00 0.01 0.02	0.01 0.01 0.02	$F_{1,225} = 0.96, P = 0.33^1$
zerofullbath80	227	0.01 0.02 0.03	0.01 0.02 0.04	0.01 0.02 0.04	$F_{1,225} = 2.1, P = 0.15^1$
northeast : 0	227	61% ( 55)	58% ( 79)	59% (134)	$\chi_1^2 = 0.27, P = 0.6^2$
1		39% ( 35)	42% ( 58)	41% ( 93)	
midwest : 0	227	68% ( 61)	72% ( 98)	70% (159)	$\chi_1^2 = 0.37, P = 0.55^2$
1		32% ( 29)	28% ( 39)	30% ( 68)	
south : 0	227	81% ( 73)	80% (109)	80% (182)	$\chi_1^2 = 0.08, P = 0.78^2$
1		19% ( 17)	20% ( 28)	20% ( 45)	
west : 0	227	90% ( 81)	91% (125)	91% (206)	$\chi_1^2 = 0.1, P = 0.75^2$
1		10% ( 9)	9% ( 12)	9% ( 21)	
meanhs8	227	33651 44351 55707	34417 46152 61835	34115 45384 59721	$F_{1,225} = 0.82, P = 0.36^1$
bedrms02_80	227	0.37 0.45 0.54	0.33 0.44 0.53	0.35 0.44 0.54	$F_{1,225} = 1.3, P = 0.25^1$
bedrms34_80	227	0.44 0.53 0.59	0.44 0.54 0.63	0.44 0.53 0.62	$F_{1,225} = 0.83, P = 0.36^1$
detach80	179	0.61 0.74 0.84	0.57 0.76 0.85	0.57 0.75 0.85	$F_{1,177} = 0.09, P = 0.76^1$
bedrms0_80occ	227	0 0 0	0 0 0	0 0 0	$F_{1,225} = 0.03, P = 0.86^1$
bedrms1_80occ	227	0.02 0.03 0.06	0.02 0.03 0.05	0.02 0.03 0.06	$F_{1,225} = 1, P = 0.31^1$
bedrms2_80occ	227	0.22 0.30 0.39	0.19 0.27 0.34	0.20 0.29 0.36	$F_{1,225} = 4.2, P = 0.041^1$
bedrms3_80occ	227	0.41 0.47 0.53	0.43 0.50 0.55	0.43 0.48 0.54	$F_{1,225} = 1.5, P = 0.21^1$
bedrms4_80occ	227	0.10 0.14 0.18	0.09 0.15 0.21	0.10 0.14 0.20	$F_{1,225} = 2.4, P = 0.12^1$
bedrms5_80occ	227	0.01 0.02 0.04	0.01 0.03 0.05	0.01 0.03 0.04	$F_{1,225} = 0.68, P = 0.41^1$
blt0_1yrs80occ	227	0.01 0.02 0.04	0.01 0.02 0.04	0.01 0.02 0.04	$F_{1,225} = 0.18, P = 0.68^1$
blt2_5yrs80occ	227	0.02 0.09 0.16	0.02 0.09 0.15	0.02 0.09 0.15	$F_{1,225} = 0.05, P = 0.83^1$
blt6_10yrs80occ	227	0.06 0.12 0.17	0.05 0.11 0.18	0.05 0.11 0.18	$F_{1,225} = 0.17, P = 0.68^1$
blt10_20yrs80occ	227	0.12 0.18 0.24	0.14 0.19 0.24	0.13 0.19 0.24	$F_{1,225} = 0.76, P = 0.38^1$
blt20_30yrs80occ	227	0.10 0.16 0.26	0.12 0.17 0.25	0.11 0.16 0.25	$F_{1,225} = 0.15, P = 0.7^1$
blt30_40yrs80occ	227	0.05 0.08 0.13	0.06 0.08 0.12	0.05 0.08 0.12	$F_{1,225} = 0.01, P = 0.94^1$
blt40_yrs80occ	227	0.11 0.19 0.33	0.11 0.19 0.34	0.11 0.19 0.34	$F_{1,225} = 0.04, P = 0.85^1$
detach80occ	227	0.82 0.91 0.97	0.85 0.93 0.98	0.83 0.92 0.97	$F_{1,225} = 2.8, P = 0.098^1$
attach80occ	227	0.00 0.00 0.01	0.00 0.00 0.01	0.00 0.00 0.01	$F_{1,225} = 0.05, P = 0.82^1$
mobile80occ	227	0.00 0.07 0.13	0.00 0.03 0.12	0.00 0.04 0.13	$F_{1,225} = 1.5, P = 0.23^1$
occupied80	227	0.93 0.95 0.97	0.92 0.95 0.96	0.92 0.95 0.96	$F_{1,225} = 0.09, P = 0.77^1$
bltmore30_80	227	0.21 0.34 0.51	0.21 0.34 0.49	0.21 0.34 0.49	$F_{1,225} = 0.11, P = 0.74^1$
og82list : 1	227	100% ( 90)	100% (137)	100% (227)	<sup>2</sup>
0		0% ( 0)	0% ( 0)	0% ( 0)	
nbr_dummy : 0	227	76% ( 68)	71% ( 97)	73% (165)	$\chi_1^2 = 0.62, P = 0.43^2$
1		24% ( 22)	29% ( 40)	27% ( 62)	

*a b c* represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables.

*N* is the number of non-missing values.

Numbers after percents are frequencies.

Tests used:

a sharp RDD. The “facts” presented in the assignment tend to support the choice. HRS was selected in rather capricious ways (“the 28.5 cutoff was selected...(for) a manageable number of sites”), was expected to be unknown to the people who generated the data, and is an imprecise (“imperfect”) scoring indicator.

## (ii) McCrary Test

It does not appear that any manipulation occurred around the threshold for listing on the NPL. Based on the plot of the density distribution (Figure 1) using default values for bandwidth and bin width ( $h \approx 12.6$ ,  $b \approx 1.6$ ), there is no significant discontinuity in the neighborhood of HRS=28.5. The estimated value lines appear to nearly match with each other, and more importantly the upperbound 95% C.I. for values below HRS=28.5 and the estimated value above HRS=28.5 overlap, and vice versa for the lowerbound 95% C.I. for values above HRS = 28.5. This indicates the choice of HRS as a running variable in the RDD is appropriate. This lack of discontinuity appears to be consistent across various bandwidth values (tested with values  $h = \{4, 8, 16, 20\}$ ).

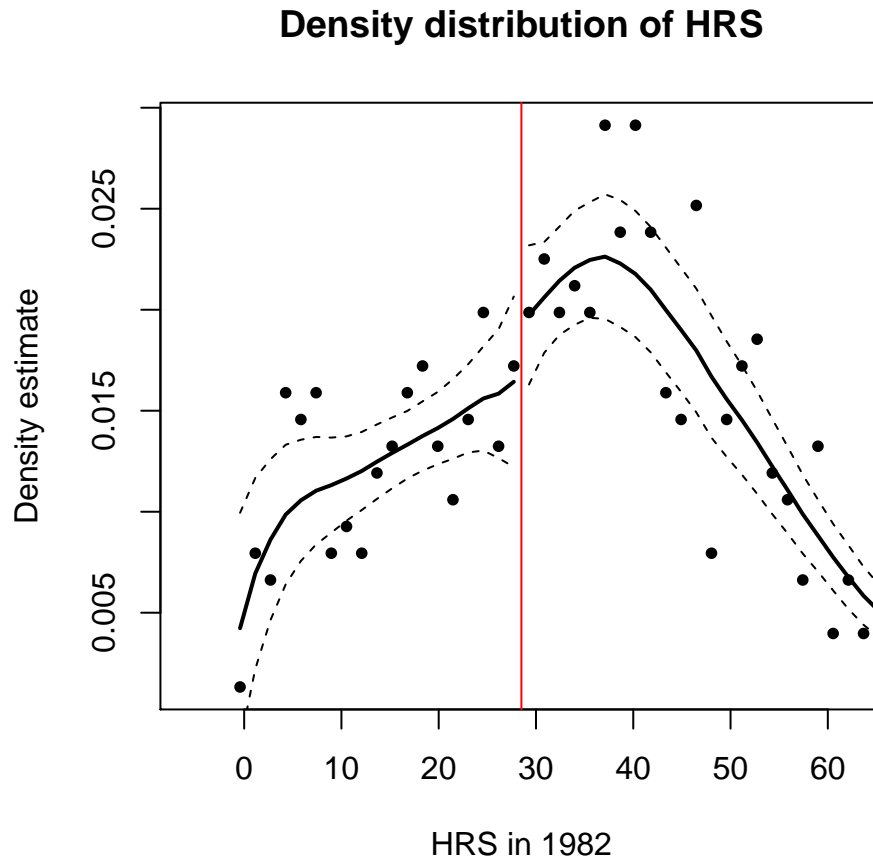


Figure 1: Density distribution histogram for 1982 HRS.

## Part C: RDD First Stage

### (i) 2SLS first stage specification

The first stage equation is:

$$NPL_{2000} = \gamma_1 \mathbf{1}\{HRS_{82} > 28.5\} + \gamma_2(HRS_{82} - c) + \gamma_3(HRS_{82} - c) * \mathbf{1}\{HRS_{82} > 28.5\} + x_i\gamma_4 + u_i.$$

We use the full set of covariates ( $x_i$ ) but not state level fixed effects in the analyses below. Both the full dataset and a constrained linear regression around the threshold (plus or minus 12 points) show a statistically significant first stage. Combined with the graphical analysis below we are confident that the first stage is useful for a fuzzy RD design.

## (ii) Graphic: NPL and HRS scores

Figure 2 shows how presence on the NPL by 2000 depends very strongly, nearly “sharply” on the value of the HRS score in 1982. Only a handful of sites do not follow the strict cutoff rule. As one would expect there are many more (but not exclusively) non-compliers with the strict boundary for cleanup on the low side of the cutoff, i.e., sites where the HRS score should not have resulted in listing. Either a revised HRS score later than 1982 or some other change (including manipulation of the process, etc.) is responsible for these cases.

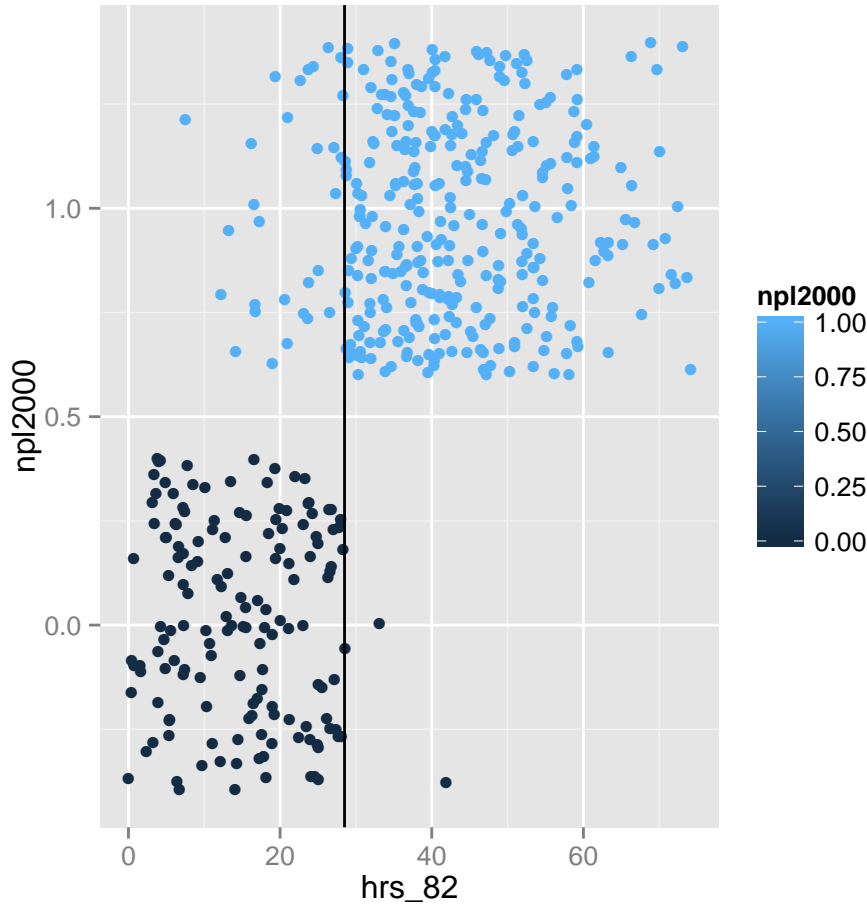


Figure 2: HRS score and status on NPL in 2000. The points are jittered to indicate density; note that the actual values are either 1 or 0.

## (iii) Placebo Test

Figure 3 shows the influence of HRS on household values in 1980. We would not expect any discontinuous jumps in this function because there was no cleanup initiated at that time. This placebo test indicates there are not endogenous discontinuities in the value of households along the HRS scale.

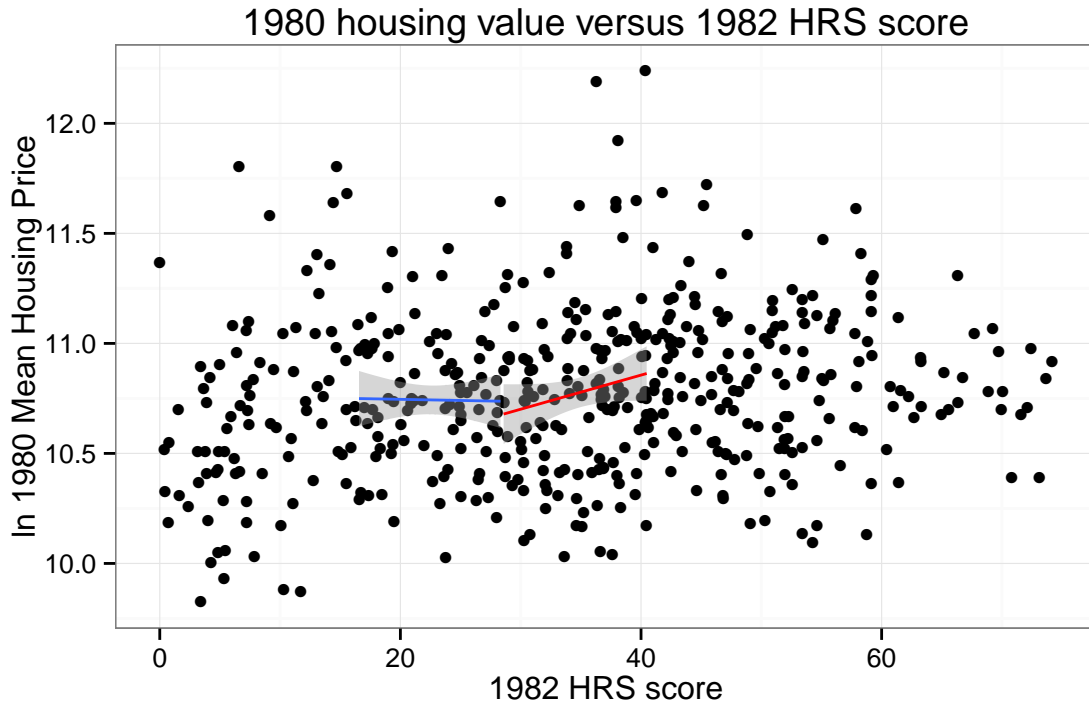


Figure 3: HRS score and 1980 mean property values. There are local fit linear regressions on either side of the 28.5 cutoff in HRS.

## Part D: RDD Second Stage

We use the reduced form for the second stage of IV estimation. In this case, it is given by:

$$\ln mdhsval_{0nbr} = \gamma_1 \mathbf{1}\{HRS_{82} > 28.5\} + \gamma_2 (HRS_{82} - c) + \gamma_3 (HRS_{82} - c) * \mathbf{1}\{HRS_{82} > 28.5\} + x_i \gamma_4 + u_i.$$

For the IV to be valid there are two assumptions that need to be met: 1) the instrument needs to be correlated with treatment, which was shown to be true in the first stage equations above, and 2) the instrument should not be correlated with the error term, which is a reasonable assumption in this case based on the placebo test analysis above.

While the first stage IV specification indicates a significant relationship between crossing the threshold and treatment status (both graphically and in the linear model formulation), the reduced form fails to show a relationship. We also used a plot more like the ones described in the notes (with local average values in evenly spaced bins and linear fits on either side of the threshold). This is shown in figure 5 below. The result is the same as the plot with the full dataset, that there is not a discernible discontinuity in the data around the threshold. The second stage fails to show any causal link between hazardous waste cleanup and housing values in the surrounding area, in spite of good specification of an instrument and a very strong first stage.

## Part E: Synthesis

An ordinary least squares setup for estimating the effect of hazardous waste cleanup on housing prices indicates we should expect that sites with National Priorities List status result in higher home values than those not on the list—an increase in value from cleanup or the potential to have cleanup. This approach is weak because it requires us to assume that we have a full set of observations for understanding the drivers of housing prices. There also appear to be issues with overlap: there are significant differences between



Table 5: 2SLS RDD of HRS@28.5 threshold vs. 2000 Housing value (constrained  $16.5 < \text{HRS} < 40.5$ )

	npl2000 First Stage	lnmdvalhs0_nbr Reduced Form
	(1)	(2)
I(hrs.82 >= 28.5)	0.626*** (0.085)	0.056 (0.066)
diff.cutoff	0.017* (0.009)	-0.002 (0.007)
tothsun8_nbr	-0.00001 (0.00001)	-0.00000 (0.00001)
occupied80_nbr	-1.461 (0.993)	3.176*** (0.775)
pop_den8_nbr	0.00000 (0.00002)	0.00001 (0.00002)
no_hs_diploma8_nbr	-0.502 (0.512)	-0.371 (0.399)
ba_or_better8_nbr	-0.238 (0.653)	-0.037 (0.510)
shrbk8_nbr	-0.101 (0.318)	-0.061 (0.248)
shrhsp8_nbr	0.295 (0.429)	0.445 (0.335)
child8_nbr	-1.396 (0.934)	-1.810** (0.728)
old8_nbr	-1.106 (0.825)	-1.138* (0.644)
shrfor8_nbr	-0.219 (0.950)	1.775** (0.741)
ffh8_nbr	1.018 (0.786)	-0.621 (0.613)
smhse8_nbr	1.160*** (0.440)	-0.013 (0.344)
hsdrop8_nbr	-0.282 (0.475)	0.004 (0.371)
unemp8_nbr	0.461 (0.956)	-1.928** (0.746)
povrat8_nbr	1.000 (0.986)	0.250 (0.769)
welfare8_nbr	0.128 (1.232)	2.251** (0.961)
avhhin8_nbr	0.00001 (0.00001)	0.00004*** (0.00001)
zerofullbath80_nbr	0.019 (2.640)	-0.600 (2.059)
firestoveheat80_nbr	0.137 (0.531)	0.461 (0.414)
nofullkitchen80_nbr	-3.906 (2.900)	1.247 (2.262)
noaircond80_nbr	0.210* (0.119)	0.223** (0.093)
ownocc8_nbr	0.00001 (0.00002)	0.00000 (0.00001)
bedrms0_80occ_nbr	-348,599.100 (1,074,601.000)	784,093.800 (838,107.300)
bedrms1_80occ_nbr	-348,594.600 (1,074,601.000)	784,108.900 (838,107.300)
bedrms2_80occ_nbr	-348,593.600 (1,074,601.000)	784,106.200 (838,107.300)
bedrms3_80occ_nbr	-348,594.100 (1,074,601.000)	784,106.300 (838,107.300)
bedrms4_80occ_nbr	-348,594.600 (1,074,601.000)	784,106.300 (838,107.300)
bedrms5_80occ_nbr	-348,593.900 (1,074,601.000)	784,109.100 (838,107.300)
detach80occ_nbr	146,018.100 (1,126,725.000)	-497,201.500 (878,759.800)
attach80occ_nbr	146,017.700 (1,126,725.000)	-497,201.900 (878,759.800)
mobile80occ_nbr	146,017.600 (1,126,725.000)	-497,201.200 (878,759.800)
blt0_1yrs80occ_nbr	-1.275 (1.513)	0.799 (1.180)
blt2_5yrs80occ_nbr	1.498** (0.718)	0.500 (0.560)
blt6_10yrs80occ_nbr	0.262 (0.562)	0.314 (0.438)
blt10_20yrs80occ_nbr	0.339 (0.400)	0.055 (0.312)
blt20_30yrs80occ_nbr	-0.251 (0.322)	0.079 (0.251)
blt30_40yrs80occ_nbr	-0.285 (0.622)	-1.121** (0.485)
I(hrs.82 >= 28.5)TRUE:diff.cutoff	-0.017 (0.012)	-0.004 (0.009)
Constant	202,577.200 (1,611,035.000)	-286,896.600 (1,256,485.000)
N	226	226
R <sup>2</sup>	0.661	0.759
Adjusted R <sup>2</sup>	0.588	0.707
Residual Std. Error (df = 185)	0.294	0.229
F Statistic (df = 40; 185)	9.029***	14.563***

Notes:

\*\*\*Significant at the 1 percent level.

\*\*Significant at the 5 percent level.

\*Significant at the 10 percent level.

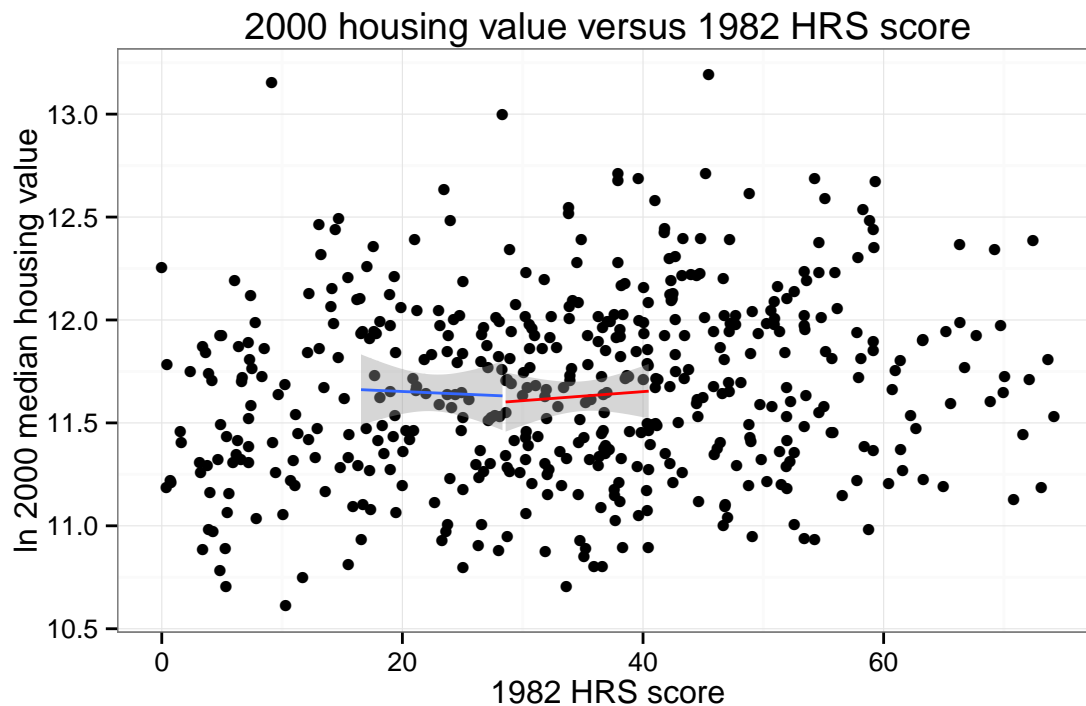


Figure 4: HRS score and 2000 median property values. There are local fit linear regressions on either side of the 28.5 cutoff in HRS.

households in areas on NPL vs. those not on the list. These differences seem to be less prominent as we narrow down on households close to the threshold for listing, motivating a regression discontinuity design. By implementing a regression discontinuity design on the same data we can test whether this potential relationship holds up around the boundary of just-under or just-over the threshold for cleanup. A well-structured RDD shows there is no relationship in the second stage despite a very strong first stage and no findings of manipulation. In other words, the threshold for listing was as good as randomly assigned near the threshold, and does not have a noticeable effect on housing prices. We trust the RDD more than OLS in this case because it is more robust in the face of error from unobserved housing characteristics and avoids overlap issues. The findings from the OLS specification may be random error or some other unobserved characteristic manifesting as a higher value from NPL status. Our overall finding is there is not apparent housing price benefit or detriment from NPL status.

A potential confounding factor that is not addressed by either research design is the stigma that may be attached to “superfund” sites. By gaining NPL status, a site has access to cleanup money but also becomes known as a superfund site instead of just a “normal” hazardous waste site. For some people who may not trust the efficacy of cleanup efforts, they could exhibit a perverse aversion to superfund sites that are under cleanup to sites that just barely missed being labeled as superfund.

## Part F: Appendix: Code Listings

```
1 ## Frank's wd
2 ## setwd("/media/frank/Data/documents/school/berkeley/fall13/are213/are213/ps3")
3 ## Peter's wd
4 # setwd("~/Google Drive/ERG/Classes/ARE213/are213/ps3")
5
6 library(foreign) #this is to read in Stata data
7 library(Hmisc)
```

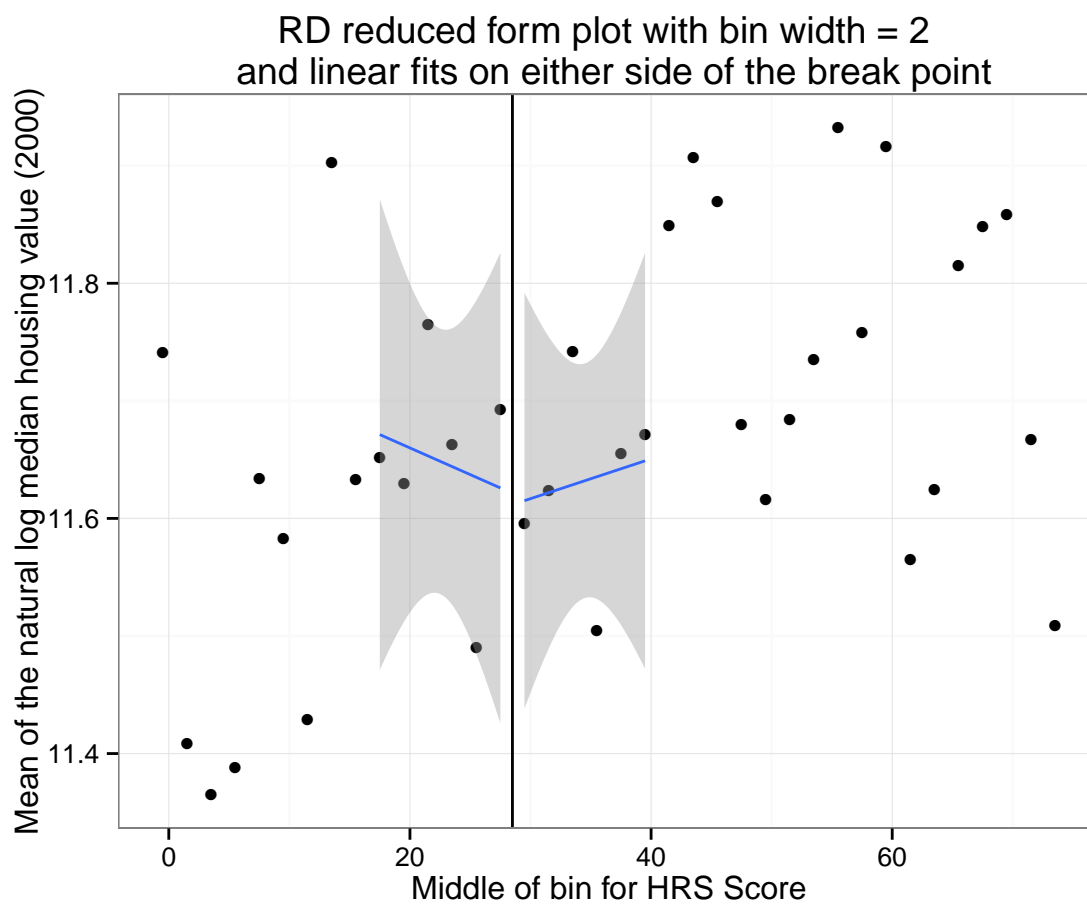


Figure 5: HRS score and 2000 median property values reported as the average in evenly spaced bins. There are local fit linear regressions on either side of the 28.5 cutoff in HRS.

```

8 library(psych)
9 library(stargazer)
10 library(ggplot2) # for neato plotting tools
11 library(plyr) # for nice data tools like ddply
12 library(car) # "companion for applied regression" - recode fxn, etc.
13 library(gmodels) #for Crosstabs
14 library(reshape)
15 library(rdd) # for regression discontinuity designs
16 library(AER) # includes ivreg function, etc. Maybe worth using...
17
18 source("../util/are213-func.R")
19
20 all.sites <- read.dta('allsites.dta')
21 all.cov <- read.dta('allcovariates.dta')
22 site.cov <- read.dta('sitecovariates.dta')
23 two.mile <- read.dta('2miledata.dta')
24
25 datakey <- function(stata.data.frame){
26   out <- data.frame(names = names(stata.data.frame),
27                     labels = attr(stata.data.frame, "var.labels"))
28   return(out)
29 }
30
31 all.sites.key <- datakey(all.sites)
32 all.cov.key <- datakey(all.cov)
33 site.cov.key <- datakey(site.cov)
34 two.mile.key <- datakey(two.mile)
35
36
37 ## Problem 1
38
39 ## Part 1a -----
40
41 reg1a.1 <- lm(data = all.sites, lnmdvalhs0 ~ npl2000 + lnmeanhs8)
42
43 # robust standard errors test
44 reg1a.1.rse <- coeftest(reg1a.1, vcov = vcovHC(reg1a.1, type="HC1"))
45
46 reg1a.2 <- lm(data = all.sites, lnmdvalhs0 ~
47               npl2000 +
48               lnmeanhs8 +
49               firestoveheat80 +
50               nofullkitchen80 +
51               zerofullbath80 +
52               ## bedrms0_80occ +
53               bedrms1_80occ +
54               bedrms2_80occ +
55               bedrms3_80occ +
56               bedrms4_80occ +
57               bedrms5_80occ +
58               blt0_1yrs80occ +
59               blt2_5yrs80occ +
60               blt6_10yrs80occ +
61               blt10_20yrs80occ +
62               blt20_30yrs80occ +
63               blt30_40yrs80occ +
64               ## blt40_yrs80occ +
65               ## detach80occ +
66               ## attach80occ +
67               ## mobile80occ +
68               occupied80)
69
70 #robust SE
71 reg1a.2.rse <- coeftest(reg1a.2, vcov = vcovHC(reg1a.2, type = "HC1"))
72
73 reg1a.3 <- lm(data = all.sites, lnmdvalhs0 ~
74               npl2000 +
75               lnmeanhs8 +
76               firestoveheat80 +
77               nofullkitchen80 +
78               zerofullbath80 +
79               ## bedrms0_80occ +
80               bedrms1_80occ +
81               bedrms2_80occ +
82               bedrms3_80occ +

```

```

83         bedrms4_80occ +
84         bedrms5_80occ +
85         blt0_1yrs80occ +
86         blt2_5yrs80occ +
87         blt6_10yrs80occ +
88         blt10_20yrs80occ +
89         blt20_30yrs80occ +
90         blt30_40yrs80occ +
91         ## blt40_yrs80occ +
92         ## detach80occ +
93         ## attach80occ +
94         ## mobile80occ +
95         occupied80 +
96         pop_den8 +
97         shrblk8 +
98         shrhsp8 +
99         child8 +
100        old8 +
101        shrfor8 +
102        ffh8 +
103        smhse8 +
104        hsdrop8 +
105        no_hs_diploma8 +
106        ba_or_better8 +
107        unemp8 +
108        povrat8 +
109        welfare8 +
110        avh8)
111
112 reg1a.3.rse <- coeftest(reg1a.3, vcov = vcovHC(reg1a.3, type = "HC1"))
113
114 reg1a.4 <- lm(data = all.sites, lnmdvalhs0 ~
115             npl2000 +
116             lnmeanhs8 +
117             firestoveheat80 +
118             nofullkitchen80 +
119             zerofullbath80 +
120             ## bedrms0_80occ +
121             bedrms1_80occ +
122             bedrms2_80occ +
123             bedrms3_80occ +
124             bedrms4_80occ +
125             bedrms5_80occ +
126             blt0_1yrs80occ +
127             blt2_5yrs80occ +
128             blt6_10yrs80occ +
129             blt10_20yrs80occ +
130             blt20_30yrs80occ +
131             blt30_40yrs80occ +
132             ## blt40_yrs80occ +
133             ## detach80occ +
134             ## attach80occ +
135             ## mobile80occ +
136             occupied80 +
137             pop_den8 +
138             shrblk8 +
139             shrhsp8 +
140             child8 +
141             old8 +
142             shrfor8 +
143             ffh8 +
144             smhse8 +
145             hsdrop8 +
146             no_hs_diploma8 +
147             ba_or_better8 +
148             unemp8 +
149             povrat8 +
150             welfare8 +
151             avh8 +
152             as.factor(statefips))
153
154 reg1a.4.rse <- coeftest(reg1a.4, vcov = vcovHC(reg1a.4, type = "HC1"))
155
156 # list of state names for fixed effects model (to omit in output)
157 state.list.omit <- vector("list", 50)

```

```

158 for(i in 1:50){state.list.omit[i] <- paste0("as.factor(statefips)",i)}
159
160 stargazer(reg1a.1.rse, reg1a.2.rse, reg1a.3.rse, reg1a.4.rse,
161           title = "Linear models for effect of NPL(2000) on housing value (with many additional state
              fixed effects omitted)",
162           column.labels = c("simple model", "+housing char.", "+demographics", "+state fixed effects"),
163           out = "tab1a.tex",
164           type = "latex",
165           style = "qje",
166           no.space = TRUE,
167           font.size = "scriptsize",
168           single.row = TRUE,
169           omit = 45:100) #omit all past first 45 vars.
170
171
172 ## Part 1b -----
173
174 # comparison of all.cov by nbr_dummy
175 all.cov.cov <- paste("npl1990+", tail(names(all.cov),-2)) #list of all covariates to use
176 formula1 <- as.formula(paste("npl2000 ~", paste(all.cov.cov, collapse="+")))
177
178 latex(summary( formula1,
179               data=all.cov,
180               method="reverse",
181               overall=TRUE,
182               long=TRUE,
183               test = TRUE
184             ),
185         title = "tab1b-1",
186         label = "tab:1b",
187         digits = 2,
188         round = 2,
189         size = "scriptsize",
190         caption = "Contingency table for a range of factors by npl2000 status",
191         exclude1=F
192       )
193
194 # over / under 28.5
195
196 site.cov$over.limit <- site.cov$hrs_82 > 28.5
197
198 site.cov.cov <- head(names(site.cov),-1) #list of all covariates to use
199 formula2 <- as.formula(paste("over.limit ~", paste(site.cov.cov, collapse="+")))
200
201 latex(summary( formula2,
202               data=site.cov,
203               method="reverse",
204               overall=TRUE,
205               long=TRUE,
206               test = TRUE
207             ),
208         title = "tab1b-2",
209         label = "tab:1b-2",
210         digits = 2,
211         round = 2,
212         size = "scriptsize",
213         caption = "Contingency table by HRS test status (over/under 28.5)",
214         exclude1=F
215       )
216
217 # narrow in on tighter window.
218
219 near.limit <- which(site.cov$hrs_82 > 16.5 & site.cov$hrs_82 < 40.5)
220
221 latex(summary( formula2,
222               data=site.cov[near.limit,],
223               method="reverse",
224               overall=TRUE,
225               long=TRUE,
226               test = TRUE
227             ),
228         title = "tab1b-3",
229         label = "tab:1b-3",
230         digits = 2,
231         round = 2,

```

```

232     size = "scriptsize",
233     caption = "Contingency table by HRS test status (JUST over/under 28.5)",
234     exclude1=F
235 )
236
237
238 ## Problem 2
239 ## Part 2a -----
240 ## Part 2b -----
241 ## Frank: I believe that what he is going for here is the DCdensity function (i.e. the McCrary
      specification test, plotted out).
242 ## I've implmented this below.
243
244 ## lim.under <- which(two.mile$hrs_82 < 28.5)
245 ## lim.over <- which(two.mile$hrs_82 >= 28.5)
246 ## gg.2b <- ggplot(two.mile, aes(hrs_82))
247 ## # todo: finish histogram with smooth lines on either side of 28.5
248 ## gg.2b <- gg.2b + geom_histogram(binwidth=5) +
249 ##   geom_vline(aes(xintercept=28.5), color = 'red')
250
251 pdf(file = './img/ddplot.pdf', width = 5, height = 5)
252 DCdensity(two.mile$hrs_82, cutpoint = 28.5)
253 abline(v = 28.5, col = "red")
254 title(main = 'Density distribution of HRS', xlab = 'HRS in 1982', ylab = 'Density estimate')
255 dev.off()
256
257 ## Problem 3
258 ## Part 3a -----
259 #instrument is whether there is a site scoring above 28.5 on the 1982 HRS
260 # We want to regress on npl2000
261 two.mile.cov <- head(names(two.mile), -14)
262 #Decided to axe the FIPS variable- any ideas how to include this and not break anything?
263 # Did you try fixed effects for the states?
264
265 two.mile$diff.cutoff <- two.mile$hrs_82 - 28.5
266
267 # I took out npl1990 as a predictor for 2000...it basically washes out all the variation that hrs should
      explain.
268
269 # I think this is the correct form for the formula (see RD class notes page 15 for fuzzy RD)
270 formula3.stage1 <- as.formula(paste("npl2000 ~ I(hrs_82 >= 28.5) + diff.cutoff + diff.cutoff:I(hrs_82 >=
      28.5) +", paste(two.mile.cov, collapse = "+"), "- blt40_yrs80occ_nbr"))
271
272 formula3.stageRF <- as.formula(paste("lnmdvalhs0_nbr ~ I(hrs_82 >= 28.5) + diff.cutoff + diff.cutoff:I(hrs
      _82 >= 28.5) +", paste(two.mile.cov, collapse = "+"), "- blt40_yrs80occ_nbr"))
273
274 stage1.1 <- lm(formula3.stage1,
275               data = two.mile)
276
277 stage1.2 <- lm(formula3.stage1,
278               data = two.mile,
279               subset = (two.mile$hrs_82 > 16.5 &
280                       two.mile$hrs_82 < 40.5)
281               )
282
283 stage2.1 <- lm(formula3.stageRF,
284               data = two.mile)
285
286 stage2.2 <- lm(formula3.stageRF,
287               data = two.mile,
288               subset = (two.mile$hrs_82 > 16.5 &
289                       two.mile$hrs_82 < 40.5)
290               )
291
292 stargazer(stage1.2, stage2.2,
293           title = "2SLS RDD of HRS@28.5 threshold vs. 2000 Housing value (constrained 16.5 $<$ HRS $<$
      40.5",
294           column.labels = c("First Stage", "Reduced Form"),
295           out = "tab3a.tex",
296           type = "latex",
297           style = "qje",
298           no.space = TRUE,
299           font.size = "scriptsize",
300           single.row = TRUE,
301           omit = 45:100) #omit all past first 45 vars.

```

```

302
303
304
305 # Function to make "local average" plots.
306 localAverageRD <- function(data, x.var, y.var, x.cutoff, binwidth){
307   out <- data.frame(x = data[[x.var]])
308   xmax <- max(data[[x.var]])
309   xmin <- min(data[[x.var]])
310   min.space <- x.cutoff - xmin
311   max.space <- xmax - x.cutoff
312   min.bins <- floor(min.space / binwidth)+1
313   max.bins <- floor(max.space / binwidth)+1
314   bin.breaks <- seq((x.cutoff-min.bins*binwidth),(x.cutoff+max.bins*binwidth), binwidth)
315   bin.labels <- data.frame(begin = bin.breaks[1:length(bin.breaks)-1], end = bin.breaks[2:length(bin.
     breaks]))
316   bin.labels$label <- seq(1, (min.bins + max.bins), 1)
317   bin.labels$middle <- (bin.labels$begin + bin.labels$end) / 2
318   for(i in 1:length(out$x)){
319     binrow <- which(bin.labels$begin <= out$x[i] & bin.labels$end >= out$x[i])
320     out$bin[i] <- bin.labels$middle[binrow]
321   }
322   out$y <- data[[y.var]]
323   out.condensed <- dplyr::ddply(out, .(bin), summarize,
324     mean.y = mean(y))
325   out.condensed$above.cutoff <- I(out.condensed$bin > x.cutoff)
326
327   return(out.condensed)
328 }
329
330 pdf("fig-locavg-2000.pdf", width = 6, height = 5)
331 binwidth <- 2
332 test.locAvg <- localAverageRD(two.mile, "hrs_82", "lnmdvalhs0_nbr", 28.5, binwidth)
333 ggplot(test.locAvg, aes(bin, mean.y))+
334   geom_point() +
335   geom_vline(aes(xintercept = 28.5)) +
336   stat_smooth(data = subset(test.locAvg, bin >= 16.5 & bin <= 40.5), method = "lm", aes(factor = as.factor
     (above.cutoff))) +
337   theme_bw() +
338   xlab("Middle of bin for HRS Score") +
339   ylab("Mean of the natural log median housing value (2000)") +
340   ggtitle(paste("RD reduced form plot with bin width =",binwidth, "\n and linear fits on either side of
     the break point"))
341 dev.off()
342
343
344
345
346
347 ## Part 3b -----
348 pdf("fig-3b.pdf", width = 6, height = 4)
349 HRSNPL.plot <- ggplot(data = two.mile, aes(x = hrs_82, y = npl2000))
350 HRSNPL.plot <- HRSNPL.plot +
351   geom_jitter(aes(color=npl2000)) +
352   geom_vline(aes(xintercept=28.5))
353 HRSNPL.plot
354 dev.off()
355
356 ## Part 3c ----- Placebo test
357 pdf("fig-3c.pdf", width = 6, height = 4)
358 HRS80val.plot <- ggplot(data = two.mile, aes(x = hrs_82, y = lnmeanhs8_nbr))
359 HRS80val.plot <- HRS80val.plot +
360   geom_point() +
361   theme_bw()+
362   stat_smooth(method = "lm", data = subset(two.mile, hrs_82 < 28.5 & hrs_82 > 16.5)) +
363   stat_smooth(method = "lm", data = subset(two.mile, hrs_82 >= 28.5 & hrs_82 < 40.5), color = "red") +
364   labs(title = "1980 housing value versus 1982 HRS score", x = "1982 HRS score", y = "ln 1980 Mean
     Housing Price")
365 HRS80val.plot
366 dev.off()
367
368
369 ## Problem 4 ---
370
371 pdf("fig-4.pdf", width = 6, height = 4)
372 HRS0val.plot <- ggplot(data = two.mile, aes(x = hrs_82, y = lnmdvalhs0_nbr))

```



```

373 HRS0val.plot <- HRS0val.plot +
374   geom_point() +
375   theme_bw() +
376   stat_smooth(method = "lm", data = subset(two.mile, hrs_82 < 28.5 & hrs_82 > 16.5)) +
377   stat_smooth(method = "lm", data = subset(two.mile, hrs_82 >= 28.5 & hrs_82 < 40.5), color = "red") +
378   labs(title = "2000 housing value versus 1982 HRS score", x = "1982 HRS score", y = "ln 2000 median
      housing value")
379 HRS0val.plot
380 dev.off()
381
382 # Attempt with canned function -----
383 k <- 1
384 i <- 20
385 formulaCAN.stageRF <- as.formula(paste("lnmdvalhs0_nbr ~ ", "hrs_82+ npl2000 | ", paste(two.mile.cov[k:i],
      collapse = "+")))
386 rd.subset <- two.mile$hrs_82 > 16.5 & two.mile$hrs_82 < 40.5
387 # doesn't seem to work with full set of covariates breaks after...why?
388 my.rdd <- RDestimate(formulaCAN.stageRF, data = two.mile, subset = rd.subset, cutpoint = 28.5, se.type = "
      HC1")
389 plot(my.rdd)

```

ps3.R

```

1 # Econometrics helper functions for [R]
2 #
3 # Peter Alstone and Frank Proulx
4 # 2013
5 # version 1
6 # contact: peter.alstone AT gmail.com
7
8 # Category: Data Management -----
9
10
11 # Category: Data Analysis -----
12
13 # Function: Find adjusted R^2 for subset of data
14 # This requires a completed linear model...pull out the relevant y-values and residuals and feed them to
      function
15 # [TODO @Peter] Improve function so it can simply evaluate lm or glm object, add error handling, general
      clean up.
16 adjr2 <- function(y,resid){
17   r2 <- 1-sum(resid^2) / sum((y-mean(y))^2)
18   return(r2)
19 } #end adjr2
20
21
22 # Category: Plots and Graphics -----
23
24 ## Function for arranging ggplots. use png(); arrange(p1, p2, ncol=1); dev.off() to save.
25 require(grid)
26 vp.layout <- function(x, y) viewport(layout.pos.row=x, layout.pos.col=y)
27 arrange_ggplot2 <- function(..., nrow=NULL, ncol=NULL, as.table=FALSE) {
28   dots <- list(...)
29   n <- length(dots)
30   if(is.null(nrow) & is.null(ncol)) { nrow = floor(n/2) ; ncol = ceiling(n/nrow)}
31   if(is.null(nrow)) { nrow = ceiling(n/ncol)}
32   if(is.null(ncol)) { ncol = ceiling(n/nrow)}
33   ## NOTE see n2mfrow in grDevices for possible alternative
34   grid.newpage()
35   pushViewport(viewport(layout=grid.layout(nrow,ncol) ) )
36   ii.p <- 1
37   for(ii.row in seq(1, nrow)){
38     ii.table.row <- ii.row
39     if(as.table) {ii.table.row <- nrow - ii.table.row + 1}
40     for(ii.col in seq(1, ncol)){
41       ii.table <- ii.p
42       if(ii.p > n) break
43       print(dots[[ii.table]], vp=vp.layout(ii.table.row, ii.col))
44       ii.p <- ii.p + 1
45     }
46   }
47 }
48
49 robust <- function(model){ #This calculates the Huber-White Robust standard errors -- code from http://
      thetarzan.wordpress.com/2011/05/28/heteroskedasticity-robust-and-clustered-standard-errors-in-r/

```

```

50 s <- summary(model)
51 X <- model.matrix(model)
52 u2 <- residuals(model)^2
53 XDX <- 0
54
55 for(i in 1:nrow(X)) {
56   XDX <- XDX + u2[i]*X[i,]%*%t(X[i,])
57 }
58
59 # inverse(X'X)
60 XX1 <- solve(t(X)%*%X)
61
62 #Compute variance/covariance matrix
63 varcovar <- XX1 %*% XDX %*% XX1
64
65 # Degrees of freedom adjustment
66 dfc <- sqrt(nrow(X))/sqrt(nrow(X)-ncol(X))
67
68 stdh <- dfc*sqrt(diag(varcovar))
69
70 t <- model$coefficients/stdh
71 p <- 2*pnorm(-abs(t))
72 results <- cbind(model$coefficients, stdh, t, p)
73 dimnames(results) <- dimnames(s$coefficients)
74 results
75 }
76
77 ## Two functions for clustered standard errors below from: http://people.su.se/~ma/clustering.pdf -----
78
79 clx <-
80 function(fm, dfcw, cluster){
81   # R-codes (www.r-project.org) for computing
82   # clustered-standard errors. Mahmood Arai, Jan 26, 2008.
83
84   # The arguments of the function are:
85   # fitted model, cluster1 and cluster2
86   # You need to install libraries 'sandwich' and 'lmtest'
87
88   # reweighting the var-cov matrix for the within model
89   library(sandwich);library(lmtest)
90   M <- length(unique(cluster))
91   N <- length(cluster)
92   K <- fm$rank
93   dfc <- (M/(M-1))*((N-1)/(N-K))
94   uj <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
95   vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)*dfcw
96   coeftest(fm, vcovCL) }
97
98 mclx <-
99 function(fm, dfcw, cluster1, cluster2){
100   # R-codes (www.r-project.org) for computing multi-way
101   # clustered-standard errors. Mahmood Arai, Jan 26, 2008.
102   # See: Thompson (2006), Cameron, Gelbach and Miller (2006)
103   # and Petersen (2006).
104   # reweighting the var-cov matrix for the within model
105
106   # The arguments of the function are:
107   # fitted model, cluster1 and cluster2
108   # You need to install libraries 'sandwich' and 'lmtest'
109
110   library(sandwich);library(lmtest)
111   cluster12 = paste(cluster1,cluster2, sep=" ")
112   M1 <- length(unique(cluster1))
113   M2 <- length(unique(cluster2))
114   M12 <- length(unique(cluster12))
115   N <- length(cluster1)
116   K <- fm$rank
117   dfc1 <- (M1/(M1-1))*((N-1)/(N-K))
118   dfc2 <- (M2/(M2-1))*((N-1)/(N-K))
119   dfc12 <- (M12/(M12-1))*((N-1)/(N-K))
120   u1j <- apply(estfun(fm), 2, function(x) tapply(x, cluster1, sum))
121   u2j <- apply(estfun(fm), 2, function(x) tapply(x, cluster2, sum))
122   u12j <- apply(estfun(fm), 2, function(x) tapply(x, cluster12, sum))
123   vc1 <- dfc1*sandwich(fm, meat=crossprod(u1j)/N )
124   vc2 <- dfc2*sandwich(fm, meat=crossprod(u2j)/N )

```

```

125     vc12 <- dfc12*sandwich(fm, meat=crossprod(u12j)/N)
126     vcovMCL <- (vc1 + vc2 - vc12)*dfcw
127     coeftest(fm, vcovMCL)}
128
129 ## Function to compute ols standard errors , robust, clustered...
130 ## Based on http://diffuseprior.wordpress.com/2012/06/15/standard-robust-and-clustered-standard-errors-
    computed-in-r/
131 ols.hetero <- function(form, data, robust=FALSE, cluster=NULL,digits=3){
132     r1 <- lm(form, data)
133     if(length(cluster)!=0){
134         data <- na.omit(data[,c(colnames(r1$model),cluster)])
135         r1 <- lm(form, data)
136     }
137     X <- model.matrix(r1)
138     n <- dim(X)[1]
139     k <- dim(X)[2]
140     if(robust==FALSE & length(cluster)==0){
141         se <- sqrt(diag(solve(crossprod(X)) * as.numeric(crossprod(resid(r1))/(n-k))))
142         res <- cbind(coef(r1),se)
143     }
144     if(robust==TRUE){
145         u <- matrix(resid(r1))
146         meat1 <- t(X) %*% diag(diag(crossprod(t(u)))) %*% X
147         dfc <- n/(n-k)
148         se <- sqrt(dfc*diag(solve(crossprod(X)) %*% meat1 %*% solve(crossprod(X))))
149         res <- cbind(coef(r1),se)
150     }
151     if(length(cluster)!=0){
152         clus <- cbind(X,data[,cluster],resid(r1))
153         colnames(clus)[(dim(clus)[2]-1):dim(clus)[2]] <- c(cluster,"resid")
154         m <- dim(table(clus[,cluster]))
155         dfc <- (m/(m-1))*((n-1)/(n-k))
156         uclust <- apply(resid(r1)*X,2, function(x) tapply(x, clus[,cluster], sum))
157         se <- sqrt(diag(solve(crossprod(X)) %*% (t(uclust) %*% uclust) %*% solve(crossprod(X))*dfc)
158         res <- cbind(coef(r1),se)
159     }
160     res <- cbind(res,res[,1]/res[,2],(1-pnorm(abs(res[,1]/res[,2])))*2)
161     res1 <- matrix(as.numeric(sprintf(paste("%. ",paste(digits,"f",sep=""),sep=""),res)),nrow=dim(res)[1])
162     rownames(res1) <- rownames(res)
163     colnames(res1) <- c("Estimate","Std. Error","t value","Pr(>|t|)")
164     return(res1)
165 }

```

../util/are213-func.R