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 soley 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)
shrblk8			−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)
shrfor8			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)
unemprt8			−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]
unemprt8	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]
mdvalhs9	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]
mdvalhs0	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:
[1]Pearson test; [2]Wilcoxon test

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

	N	FALSE	TRUE	Combined	Test Statistic
		$N = 181$	$N = 306$	$N = 487$	
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)	$\chi^2_1 = 383,\ P < 0.001$[2]
1		13% (23)	99% (303)	67% (326)	
npl2000 : 0	487	84% (152)	1% (3)	32% (155)	$\chi^2_1 = 361,\ P < 0.001$[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]
shrblk8	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]
avhhin8	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)	$\chi^2_1 = 9.9,\ P = 0.002$[2]
1		33% (60)	48% (146)	42% (206)	
midwest : 0	487	65% (118)	77% (237)	73% (355)	$\chi^2_1 = 8.7,\ P = 0.003$[2]
1		35% (63)	23% (69)	27% (132)	
south : 0	487	78% (142)	81% (247)	80% (389)	$\chi^2_1 = 0.36,\ P = 0.55$[2]
1		22% (39)	19% (59)	20% (98)	
west : 0	487	90% (162)	90% (274)	90% (436)	$\chi^2_1 = 0,\ P = 0.99$[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)	[2]
0		0% (0)	0% (0)	0% (0)	
nbr_dummy : 0	487	80% (144)	66% (203)	71% (347)	$\chi^2_1 = 9.7,\ P = 0.002$[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 $N = 90$	TRUE $N = 137$	Combined $N = 227$	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)	$\chi^2_1 = 146,\ P < 0.001$[2]
1		22% (20)	99% (135)	68% (155)	
npl2000 : 0	227	73% (66)	1% (2)	30% (68)	$\chi^2_1 = 134,\ P < 0.001$[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]
shrblk8	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]
shrhsp8	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]
unemprt8	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)	$\chi^2_1 = 0.27,\ P = 0.6$[2]
1		39% (35)	42% (58)	41% (93)	
midwest : 0	227	68% (61)	72% (98)	70% (159)	$\chi^2_1 = 0.37,\ P = 0.55$[2]
1		32% (29)	28% (39)	30% (68)	
south : 0	227	81% (73)	80% (109)	80% (182)	$\chi^2_1 = 0.08,\ P = 0.78$[2]
1		19% (17)	20% (28)	20% (45)	
west : 0	227	90% (81)	91% (125)	91% (206)	$\chi^2_1 = 0.1,\ P = 0.75$[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)	[2]
0		0% (0)	0% (0)	0% (0)	
nbr_dummy : 0	227	76% (68)	71% (97)	73% (165)	$\chi^2_1 = 0.62,\ P = 0.43$[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 occured 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 $h = \{4, 8, 16, 20\}$).

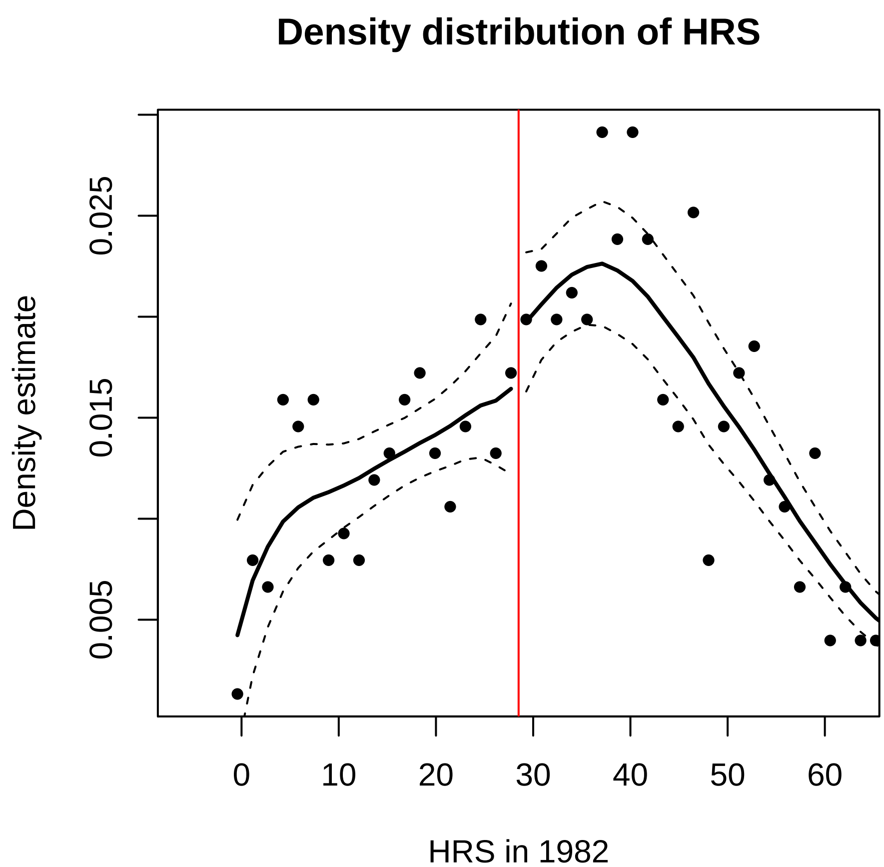

Figure 1: Density distribution histogram for 1982 HRS.

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.

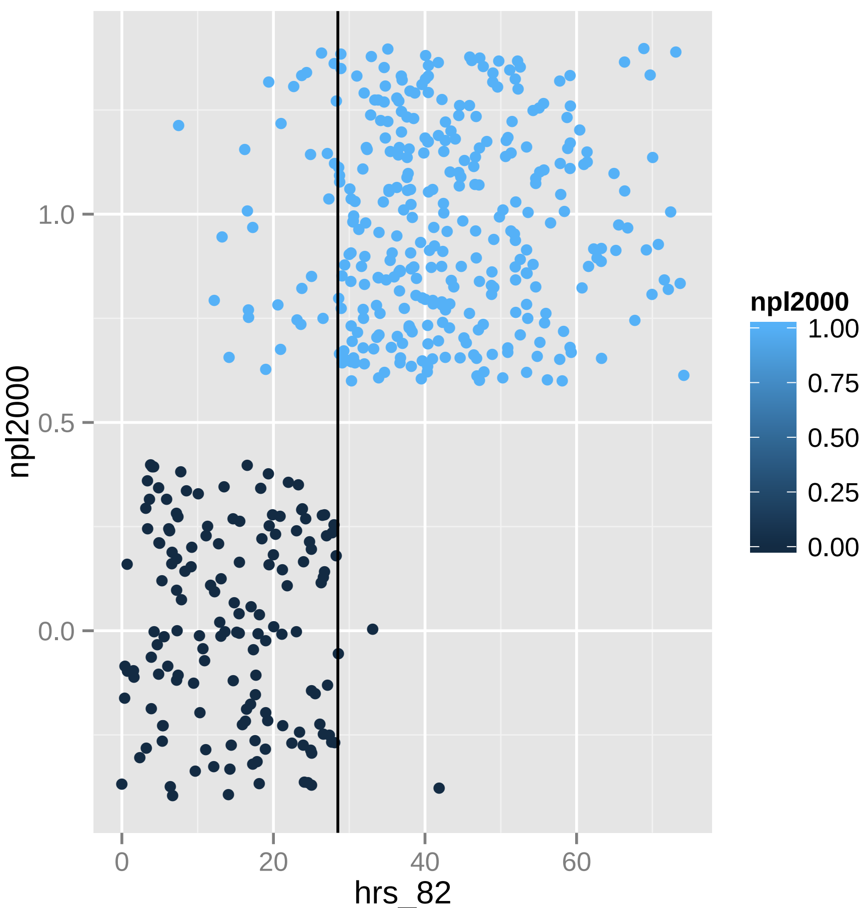

Figure 2: HRS score and status on NPL in 2000. The points are jittered to indicate density; note that the actual values are either 1 or 0.

(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.

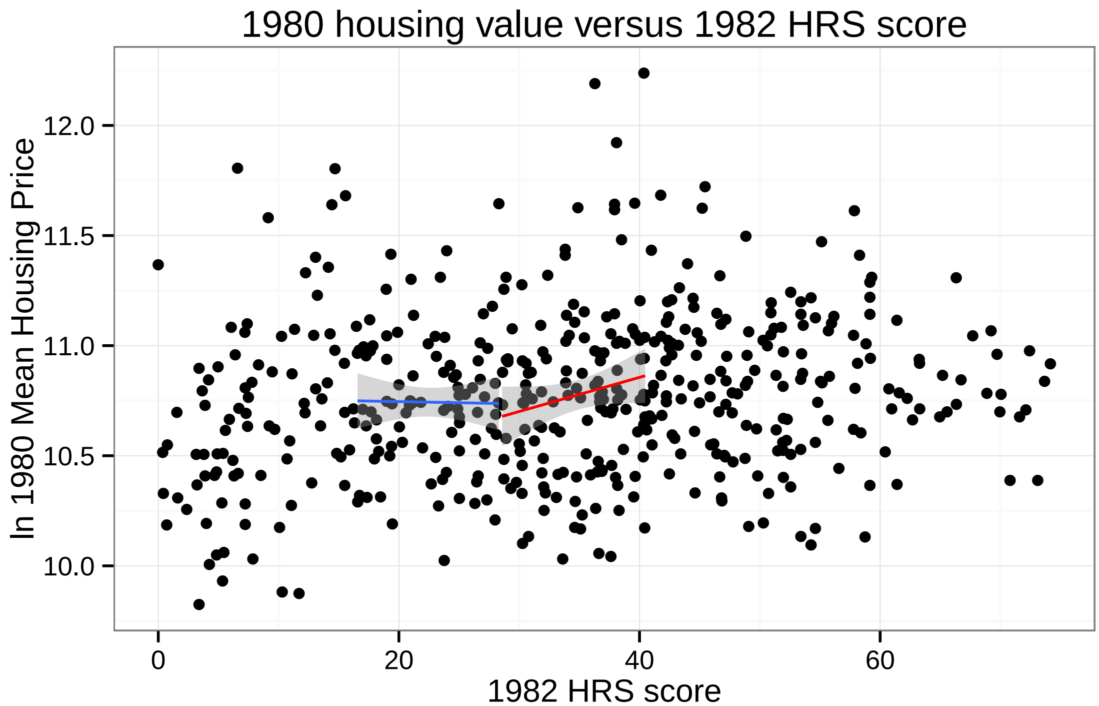

Figure 3: HRS score and 1980 mean property values. There are local fit linear regressions on either side of the 28.5 cutoff in HRS.

Part D: RDD Second Stage

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

$lnmdhsval0_{nbr} = \gamma_1 \mathbf{1}\{HRS_{82} > 28.5\} + \gamma_2(HRS_{82} - c) + \gamma_3(HRS_{82} - c) * \mathbf{1}\{HRS_{82} > 28.5\} + x_i\gamma_4 + u_i.$

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 < HRS < 40.5

	npl2000 First Stage (1)	lnmdvalhs0_nbr Reduced Form (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)
shrblk8_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)
unemprt8_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^2	0.661	0.759
Adjusted R^2	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.

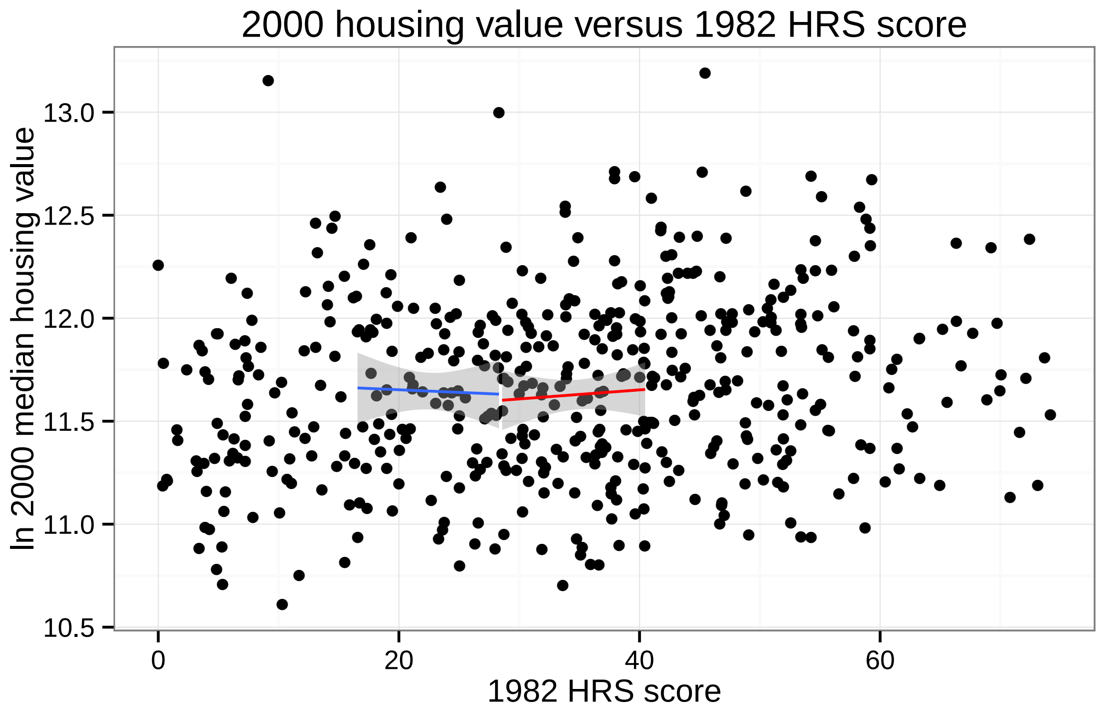

Figure 4: HRS score and 2000 median property values. There are local fit linear regressions on either side of the 28.5 cutoff in HRS.

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 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" bazardous 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

```
## Frank's wd
## setwd("/media/frank/Data/documents/school/berkeley/fall13/are213/are213/ps3")
## Peter's wd
# setwd("~/Google Drive/ERG/Classes/ARE213/are213/ps3")

library(foreign) #this is to read in Stata data
library(Hmisc)
```

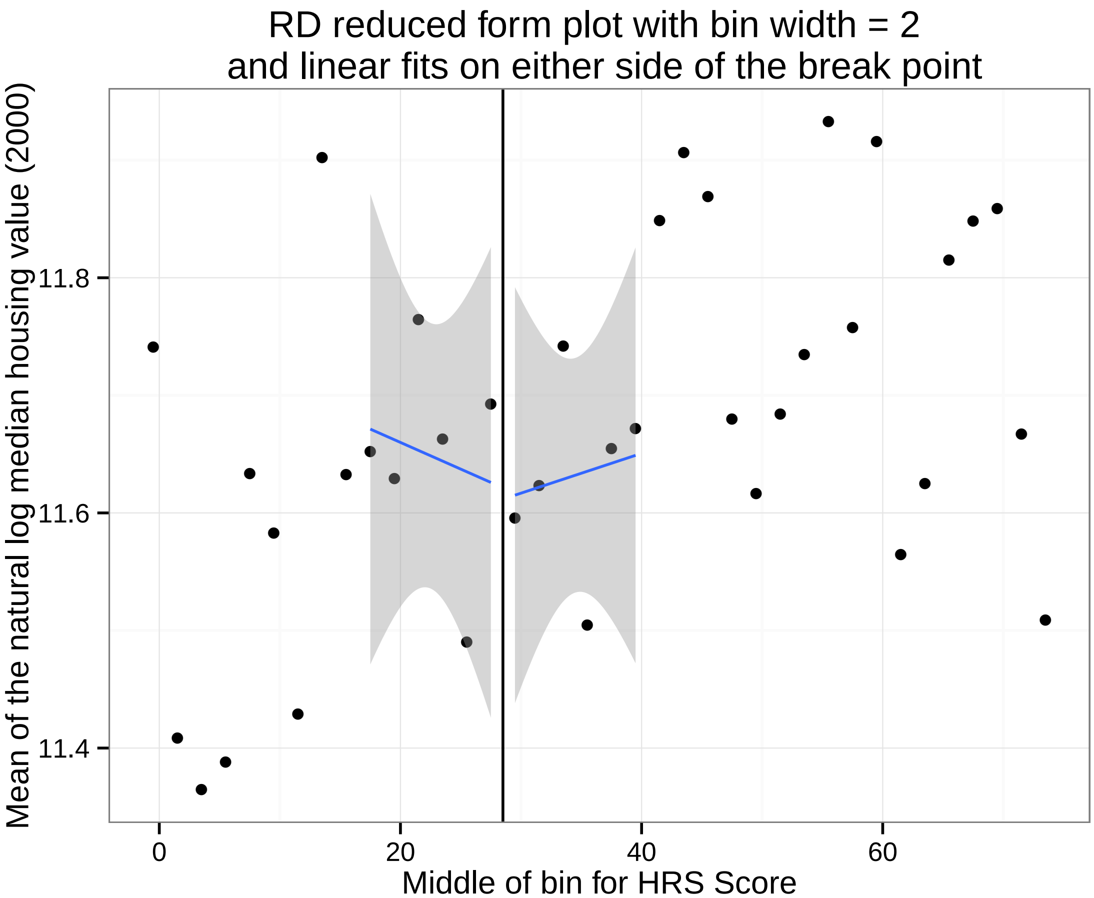

Figure 5: HRS score and 2000 median property values reported as the average in evenly spaced bins. There are local fit linear regressions on either side of the 28.5 cutoff in HRS.

```
library(psych)
library(stargazer)
library(ggplot2) # for neato plotting tools
library(plyr) # for nice data tools like ddply
library(car) # "companion for applied regression" - recode fxn, etc.
library(gmodels) #for Crosstabs
library(reshape)
library(rdd) # for regression discontinuity designs
library(AER) # includes ivreg function, etc.  Maybe worth using...

source("../util/are213-func.R")

all.sites <- read.dta('allsites.dta')
all.cov <- read.dta('allcovariates.dta')
site.cov <- read.dta('sitecovariates.dta')
two.mile <- read.dta('2miledata.dta')

datakey <- function(stata.data.frame){
  out <- data.frame(names = names(stata.data.frame),
                    labels = attr(stata.data.frame, "var.labels"))
  return(out)
}

all.sites.key <- datakey(all.sites)
all.cov.key <- datakey(all.cov)
site.cov.key <- datakey(site.cov)
two.mile.key <- datakey(two.mile)


## Problem 1

## Part 1a -------

reg1a.1 <- lm(data = all.sites, lnmdvalhs0 ~ npl2000 + lnmeanhs8)

# robust standard errors test
reg1a.1.rse <- coeftest(reg1a.1, vcov = vcovHC(reg1a.1, type="HC1"))

reg1a.2 <- lm(data = all.sites, lnmdvalhs0 ~
                  npl2000 +
                  lnmeanhs8 +
                  firestoveheat80 +
                  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)

#robust SE
reg1a.2.rse <- coeftest(reg1a.2, vcov = vcovHC(reg1a.2, type = "HC1"))

reg1a.3 <- lm(data = all.sites, lnmdvalhs0 ~
                  npl2000 +
                  lnmeanhs8 +
                  firestoveheat80 +
                  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)

reg1a.3.rse <- coeftest(reg1a.3, vcov = vcovHC(reg1a.3, type = "HC1"))

reg1a.4 <- lm(data = all.sites, lnmdvalhs0 ~
                npl2000 +
                lnmeanhs8 +
                firestoveheat80 +
                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 +
                as.factor(statefips))

reg1a.4.rse <- coeftest(reg1a.4, vcov = vcovHC(reg1a.4, type = "HC1"))

# list of state names for fixed effects model (to omit in output)
state.list.omit <- vector("list", 50)
```

```
for(i in 1:50){state.list.omit[i] <- paste0("as.factor(statefips)",i)}

stargazer(reg1a.1.rse, reg1a.2.rse, reg1a.3.rse, reg1a.4.rse,
          title = "Linear models for effect of NPL(2000) on housing value (with many additional state fixed effects omitted)",
          column.labels = c("simple model", "+housing char.", "+demographics", "+state fixed effects"),
          out = "tab1a.tex",
          type = "latex",
          style = "qje",
          no.space = TRUE,
          font.size = "scriptsize",
          single.row = TRUE,
          omit = 45:100) #omit all past first 45 vars.


## Part 1b -------

# comparison of all.cov by nbr_dummy
all.cov.cov <- paste("npl1990+", tail(names(all.cov),-2)) #list of all covariates to use
formula1 <- as.formula(paste("npl2000 ~", paste(all.cov.cov, collapse="+")))

latex(summary( formula1,
               data=all.cov,
               method="reverse",
               overall=TRUE,
               long=TRUE,
               test = TRUE
),
      title = "tab1b-1",
      label = "tab:1b",
      digits = 2,
      round = 2,
      size = "scriptsize",
      caption = "Contingency table for a range of factors by npl2000 status",
      exclude1=F
)

# over / under 28.5

site.cov$over.limit <- site.cov$hrs_82 > 28.5

site.cov.cov <- head(names(site.cov),-1) #list of all covariates to use
formula2 <- as.formula(paste("over.limit ~", paste(site.cov.cov, collapse="+")))

latex(summary( formula2,
               data=site.cov,
               method="reverse",
               overall=TRUE,
               long=TRUE,
               test = TRUE
),
      title = "tab1b-2",
      label = "tab:1b-2",
      digits = 2,
      round = 2,
      size = "scriptsize",
      caption = "Contingency table by HRS test status (over/under 28.5)",
      exclude1=F
)

# narrow in on tighter window.

near.limit <- which(site.cov$hrs_82 > 16.5 & site.cov$hrs_82 < 40.5)

latex(summary( formula2,
               data=site.cov[near.limit,],
               method="reverse",
               overall=TRUE,
               long=TRUE,
               test = TRUE
),
      title = "tab1b-3",
      label = "tab:1b-3",
      digits = 2,
      round = 2,
```

```
      size = "scriptsize",
      caption = "Contingency table by HRS test status (JUST over/under 28.5)",
      exclude1=F
)


## Problem 2
## Part 2a --------
## Part 2b -------
## Frank: I believe that what he is going for here is the DCdensity function (i.e. the McCrary specification test, plotted out).
## I've implmented this below.

## lim.under <- which(two.mile$hrs_82 < 28.5)
## lim.over <- which(two.mile$hrs_82 >= 28.5)
## gg.2b <- ggplot(two.mile, aes(hrs_82))
## # todo: finish histogram with smooth lines on either side of 28.5
## gg.2b <- gg.2b + geom_histogram(binwidth=5) +
##   geom_vline(aes(xintercept=28.5), color = 'red')

pdf(file = './img/ddplot.pdf', width = 5, height = 5)
DCdensity(two.mile$hrs_82, cutpoint = 28.5)
abline(v = 28.5, col = "red")
title(main = 'Density distribution of HRS', xlab = 'HRS in 1982', ylab = 'Density estimate')
dev.off()

## Problem 3
## Part 3a ------
#instrument is whether there is a site scoring above 28.5 on the 1982 HRS
# We want to regress on npl2000
two.mile.cov <- head(names(two.mile), -14)
#Decided to axe the FIPS variable- any ideas how to include this and not break anything?
# Did you try fixed effects for the states?

two.mile$diff.cutoff <- two.mile$hrs_82 - 28.5

# I took out npl1990 as a predictor for 2000...it basically washes out all the variation that hrs shoudl explain.

# I think this is the correct form for the formula (see RD class notes page 15 for fuzzy RD)
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"))

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"))

stage1.1 <- lm(formula3.stage1,
               data = two.mile)

stage1.2 <- lm(formula3.stage1,
               data = two.mile,
               subset = (two.mile$hrs_82 > 16.5 &
                         two.mile$hrs_82 < 40.5)
               )

stage2.1 <- lm(formula3.stageRF,
               data = two.mile)

stage2.2 <- lm(formula3.stageRF,
               data = two.mile,
               subset = (two.mile$hrs_82 > 16.5 &
                           two.mile$hrs_82 < 40.5)
)

stargazer(stage1.2, stage2.2,
          title = "2SLS RDD of HRS@28.5 threshold vs. 2000 Housing value (constrained 16.5 $<$ HRS $<$ 40.5",
          column.labels = c("First Stage", "Reduced Form"),
          out = "tab3a.tex",
          type = "latex",
          style = "qje",
          no.space = TRUE,
          font.size = "scriptsize",
          single.row = TRUE,
          omit = 45:100) #omit all past first 45 vars.
```

```r



# Function to make "local average" plots.
localAverageRD <- function(data, x.var, y.var, x.cutoff, binwidth){
  out <- data.frame(x = data[[x.var]])
  xmax <- max(data[[x.var]])
  xmin <- min(data[[x.var]])
  min.space <- x.cutoff - xmin
  max.space <- xmax - x.cutoff
  min.bins <- floor(min.space / binwidth)+1
  max.bins <- floor(max.space / binwidth)+1
  bin.breaks <- seq((x.cutoff-min.bins*binwidth),(x.cutoff+max.bins*binwidth), binwidth)
  bin.labels <- data.frame(begin = bin.breaks[1:length(bin.breaks)-1], end = bin.breaks[2:length(bin.breaks)])
  bin.labels$label <- seq(1, (min.bins + max.bins), 1)
  bin.labels$middle <- (bin.labels$begin + bin.labels$end) / 2
  for(i in 1:length(out$x)){
    binrow <- which(bin.labels$begin <= out$x[i] & bin.labels$end >= out$x[i])
    out$bin[i] <- bin.labels$middle[binrow]
  }
  out$y <- data[[y.var]]
  out.condensed <- ddply(out, .(bin), summarize,
                         mean.y = mean(y))
  out.condensed$above.cutoff <- I(out.condensed$bin > x.cutoff)

  return(out.condensed)
}

pdf("fig-locavg-2000.pdf", width = 6, height = 5)
binwidth <- 2
test.locAvg <- localAverageRD(two.mile, "hrs_82", "lnmdvalhs0_nbr", 28.5, binwidth)
ggplot(test.locAvg, aes(bin, mean.y))+
  geom_point() +
  geom_vline(aes(xintercept = 28.5)) +
  stat_smooth(data = subset(test.locAvg, bin >= 16.5 & bin <= 40.5), method = "lm", aes(factor = as.factor(above.cutoff))) +
  theme_bw() +
  xlab("Middle of bin for HRS Score") +
  ylab("Mean of the natural log median housing value (2000)") +
  ggtitle(paste("RD reduced form plot with bin width =",binwidth, "\n and linear fits on either side of the break point"))
dev.off()





## Part 3b ------
pdf("fig-3b.pdf", width = 6, height = 4)
HRSNPL.plot <- ggplot(data = two.mile, aes(x = hrs_82, y = npl2000))
HRSNPL.plot <- HRSNPL.plot +
    geom_jitter(aes(color=npl2000)) +
  geom_vline(aes(xintercept=28.5))
HRSNPL.plot
dev.off()

## Part 3c ------- Placebo test
pdf("fig-3c.pdf", width = 6, height = 4)
HRS80val.plot <- ggplot(data = two.mile, aes(x = hrs_82, y = lnmeanhs8_nbr))
HRS80val.plot <- HRS80val.plot +
    geom_point() +
  theme_bw()+
  stat_smooth(method = "lm", data = subset(two.mile, hrs_82 < 28.5 & hrs_82 > 16.5)) +
  stat_smooth(method = "lm", data = subset(two.mile, hrs_82 >= 28.5 & hrs_82 < 40.5), color = "red") +
    labs(title = "1980 housing value versus 1982 HRS score", x = "1982 HRS score", y = "ln 1980 Mean Housing Price")
HRS80val.plot
dev.off()


## Problem 4 ---

pdf("fig-4.pdf", width = 6, height = 4)
HRS0val.plot <- ggplot(data = two.mile, aes(x = hrs_82, y = lnmdvalhs0_nbr))
```

```r
HRS0val.plot <- HRS0val.plot +
  geom_point() +
  theme_bw() +
  stat_smooth(method = "lm", data = subset(two.mile, hrs_82 < 28.5 & hrs_82 > 16.5)) +
  stat_smooth(method = "lm", data = subset(two.mile, hrs_82 >= 28.5 & hrs_82 < 40.5), color = "red") +
  labs(title = "2000 housing value versus 1982 HRS score", x = "1982 HRS score", y = "ln 2000 median housing value")
HRS0val.plot
dev.off()

# Attempt with canned function -------
k <- 1
i <- 20
formulaCAN.stageRF <- as.formula(paste("lnmdvalhs0_nbr ~ ", "hrs_82+ npl2000 | ", paste(two.mile.cov[k:i], collapse = "+")))
rd.subset <- two.mile$hrs_82 > 16.5 & two.mile$hrs_82 < 40.5
# doesn't seem to work with full set of covariates breaks after...why?
my.rdd <- RDestimate(formulaCAN.stageRF, data = two.mile, subset = rd.subset, cutpoint = 28.5, se.type = "HC1")
plot(my.rdd)
```

ps3.r

```r
# Econometrics helper functions for [R]
#
# Peter Alstone and Frank Proulx
# 2013
# version 1
# contact: peter.alstone AT gmail.com

# Category: Data Management -------------


# Category: Data Analysis ----------------

# Function: Find adjusted R^2 for subset of data
# This requires a completed linear model...pull out the relevant y-values and residuals and feed them to function
# [TODO @Peter] Improve function so it can simply evaluate lm or glm object, add error handling, general clean up.
adjr2 <- function(y,resid){
  r2 <- 1-sum(resid^2) / sum((y-mean(y))^2)
  return(r2)
} #end adjr2


# Category: Plots and Graphics -------------

## Function for arranging ggplots. use png(); arrange(p1, p2, ncol=1); dev.off() to save.
require(grid)
vp.layout <- function(x, y) viewport(layout.pos.row=x, layout.pos.col=y)
arrange_ggplot2 <- function(..., nrow=NULL, ncol=NULL, as.table=FALSE) {
  dots <- list(...)
  n <- length(dots)
  if(is.null(nrow) & is.null(ncol)) { nrow = floor(n/2) ; ncol = ceiling(n/nrow)}
  if(is.null(nrow)) { nrow = ceiling(n/ncol)}
  if(is.null(ncol)) { ncol = ceiling(n/nrow)}
  ## NOTE see n2mfrow in grDevices for possible alternative
  grid.newpage()
  pushViewport(viewport(layout=grid.layout(nrow,ncol) ) )
  ii.p <- 1
  for(ii.row in seq(1, nrow)){
    ii.table.row <- ii.row
    if(as.table) {ii.table.row <- nrow - ii.table.row + 1}
    for(ii.col in seq(1, ncol)){
      ii.table <- ii.p
      if(ii.p > n) break
      print(dots[[ii.table]], vp=vp.layout(ii.table.row, ii.col))
      ii.p <- ii.p + 1
    }
  }
}

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/
```

```
    s <- summary(model)
    X <- model.matrix(model)
    u2 <- residuals(model)^2
    XDX <- 0

    for(i in 1:nrow(X)) {
        XDX <- XDX +u2[i]*X[i,]%*%t(X[i,])
    }

# inverse(X'X)
    XX1 <- solve(t(X)%*%X)

#Compute variance/covariance matrix
    varcovar <- XX1 %*% XDX %*% XX1

# Degrees of freedom adjustment
    dfc <- sqrt(nrow(X))/sqrt(nrow(X)-ncol(X))

    stdh <- dfc*sqrt(diag(varcovar))

    t <- model$coefficients/stdh
    p <- 2*pnorm(-abs(t))
    results <- cbind(model$coefficients, stdh, t, p)
    dimnames(results) <- dimnames(s$coefficients)
    results
}

## Two functions for clustered standard errors below from: http://people.su.se/~ma/clustering.pdf -------

clx <-
  function(fm, dfcw, cluster){
    # R-codes (www.r-project.org) for computing
    # clustered-standard errors. Mahmood Arai, Jan 26, 2008.

    # The arguments of the function are:
    # fitted model, cluster1 and cluster2
    # You need to install libraries `sandwich' and `lmtest'

    # reweighting the var-cov matrix for the within model
    library(sandwich);library(lmtest)
    M <- length(unique(cluster))
    N <- length(cluster)
    K <- fm$rank
    dfc <- (M/(M-1))*((N-1)/(N-K))
    uj  <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
    vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)*dfcw
    coeftest(fm, vcovCL) }

mclx <-
  function(fm, dfcw, cluster1, cluster2){
    # R-codes (www.r-project.org) for computing multi-way
    # clustered-standard errors. Mahmood Arai, Jan 26, 2008.
    # See: Thompson (2006), Cameron, Gelbach and Miller (2006)
    # and Petersen (2006).
    # reweighting the var-cov matrix for the within model

    # The arguments of the function are:
    # fitted model, cluster1 and cluster2
    # You need to install libraries `sandwich' and `lmtest'

    library(sandwich);library(lmtest)
    cluster12 = paste(cluster1,cluster2, sep="")
    M1  <- length(unique(cluster1))
    M2  <- length(unique(cluster2))
    M12 <- length(unique(cluster12))
    N   <- length(cluster1)
    K   <- fm$rank
    dfc1  <- (M1/(M1-1))*((N-1)/(N-K))
    dfc2  <- (M2/(M2-1))*((N-1)/(N-K))
    dfc12 <- (M12/(M12-1))*((N-1)/(N-K))
    u1j   <- apply(estfun(fm), 2, function(x) tapply(x, cluster1,  sum))
    u2j   <- apply(estfun(fm), 2, function(x) tapply(x, cluster2,  sum))
    u12j  <- apply(estfun(fm), 2, function(x) tapply(x, cluster12, sum))
    vc1   <-  dfc1*sandwich(fm, meat=crossprod(u1j)/N )
    vc2   <-  dfc2*sandwich(fm, meat=crossprod(u2j)/N )
```

```r
    vc12  <- dfc12*sandwich(fm, meat=crossprod(u12j)/N)
    vcovMCL <- (vc1 + vc2 - vc12)*dfcw
    coeftest(fm, vcovMCL)}

## Function to compute ols standard errors , robust, clustered...
## Based on http://diffuseprior.wordpress.com/2012/06/15/standard-robust-and-clustered-standard-errors-
    computed-in-r/
ols.hetero <- function(form, data, robust=FALSE, cluster=NULL,digits=3){
  r1 <- lm(form, data)
  if(length(cluster)!=0){
    data <- na.omit(data[,c(colnames(r1$model),cluster)])
    r1 <- lm(form, data)
  }
  X <- model.matrix(r1)
  n <- dim(X)[1]
  k <- dim(X)[2]
  if(robust==FALSE & length(cluster)==0){
    se <- sqrt(diag(solve(crossprod(X)) * as.numeric(crossprod(resid(r1))/(n-k))))
    res <- cbind(coef(r1),se)
  }
  if(robust==TRUE){
    u <- matrix(resid(r1))
    meat1 <- t(X) %*% diag(diag(crossprod(t(u)))) %*% X
    dfc <- n/(n-k)
    se <- sqrt(dfc*diag(solve(crossprod(X)) %*% meat1 %*% solve(crossprod(X))))
    res <- cbind(coef(r1),se)
  }
  if(length(cluster)!=0){
    clus <- cbind(X,data[,cluster],resid(r1))
    colnames(clus)[(dim(clus)[2]-1):dim(clus)[2]] <- c(cluster,"resid")
    m <- dim(table(clus[,cluster]))
    dfc <- (m/(m-1))*((n-1)/(n-k))
    uclust  <- apply(resid(r1)*X,2, function(x) tapply(x, clus[,cluster], sum))
    se <- sqrt(diag(solve(crossprod(X)) %*% (t(uclust) %*% uclust) %*% solve(crossprod(X)))*dfc)
    res <- cbind(coef(r1),se)
  }
  res <- cbind(res,res[,1]/res[,2],(1-pnorm(abs(res[,1]/res[,2])))*2)
  res1 <- matrix(as.numeric(sprintf(paste("%.",paste(digits,"f",sep=""),sep=""),res)),nrow=dim(res)[1])
  rownames(res1) <- rownames(res)
  colnames(res1) <- c("Estimate","Std. Error","t value","Pr(>|t|)")
  return(res1)
}
```

../util/are213–func.R