1 Regression Discontinuity

In Israel, schools face a rule which states that classes cannot be larger than 40 pupils. When enrollment is 41, schools are supposed to open a second classroom, and then open a third classroom at 81 pupils etc. This causes discontinuous drops in class size at multiples of 40.

Problem 1.1. Estimate the effect of class size on math scores using OLS without any controls, and then by adding the percentage of disadvantaged students in the class and enrollment as controls. Interpret your results.

Solution. We first estimate the effect of class size on math scores using OLS without any controls:

avgmath	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
classize	.324184	.0330271	9.82	0.000	.2594134	.3889547
_cons	57.61222	1.013029	56.87	0.000	55.62553	59.59891

and then we add the percentage of disadvantaged students in the class and enrollment as controls.

avgmath	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
classize	.0230558	.0385498	0.60	0.550	-.0525456	.0986573
tipuach	-.3220914	.0158721	-20.29	0.000	-.3532188	-.2909639
c_size	.0192267	.0065736	2.92	0.003	.006335	.0321185
_cons	69.67829	1.074253	64.86	0.000	67.57154	71.78505

▷ Absent controls, an extra student raises the math score by 0.3241 points. Adding controls reduces this effect to 0.0231 points. Using OLS here will not tease out the causal effect since the estimate includes both the average effect on the treated and a selection bias term.

Start by limiting the sample to schools with enrollment between 20 and 60 students. Generate a (predicted) large class dummy based on the first discontinuity at 40 students.

Problem 1.2. Use OLS to estimate the effect of being in a large class on math scores assuming that you have a sharp RDD around this discontinuity. Control for the percentage of disadvantaged students in the class and a linear trend in enrollment.

Solution. To control for the linear trend in enrollment, we compute the number of enrolled students above 40 and also interact it with the large class dummy. Thus, the regression specification looks as the following:

$$Y = \alpha + \zeta Z + \beta\left(\text{c\_size} - 40\right) + \gamma\left(\text{c\_size } - 40\right) Z + X'\eta + \epsilon$$

where Z is the large_class dummy. Under the sharp RDD assumption , ζ represents the causal effect of having a large class on average math score.

▷ Running the regression, we then obtain the following estimates:

avgmath	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
large	-4.299815	1.405924	-3.06	0.002	-7.06019	-1.53944
c_size_over	.0183108	.0688914	0.27	0.790	-.1169498	.1535714
c_size_interact	-.2830043	.1134731	-2.49	0.013	-.5057961	-.0602126
tipuach	-.3375029	.0194353	-17.37	0.000	-.375662	-.2993438
_cons	71.81067	.9345811	76.84	0.000	69.97572	73.64561

▷ Assuming we have a sharp RDD around this continuity, we find that an additional person in class reduces the math score by 4.299815 points.

Problem 1.3. Use Local Linear Regression (use the command -lpoly- in Stata) to get a point estimate of the effect of being in a large class on math scores assuming you have a sharp RDD. Finally, use a nonparametric bootstrap to estimate the standard error on your RDD point estimate. Compare these results to the estimates you obtained with OLS.

Solution. Now instead of using OLS, we will use the local linear regression to get a point estimate of the effect, assuming we have a sharp RDD. We obtain the following point estimate:

```
. di rddest
-3.5780141
```

▷ We find that the point estimate of large class size on average math score is -3.578.

We now use non-parametric bootstrap to estimate the standard error. Note that we draw samples of size 700 with replacement from the original data when performing the bootstrap estimation.

variable	mean	se(mean)
_rddest	-3.741193	.3377325

▷ We obtain a standard error of 0.3377.

▷ In the OLS estimation, we had a standard error of 1.405 associated with an estimate of -4.299. We find that the results are quantitatively similar.

Problem 1.4. Estimate the effect of class size on math scores using fuzzy RDD. Control for the percentage of disadvantaged students in the class and a linear trend in enrollment.

Solution. Now we estimate the effect using fuzzy RDD. Essentially, we are using the probability of being treated as our instrument and run a two-stage least-squares regression:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
classize	-.3775727	.1568753	-2.41	0.016	-.6850427	-.0701027
c_size	.0602166	.0313408	1.92	0.055	-.0012102	.1216434
tipuach	-.3535195	.0213088	-16.59	0.000	-.3952839	-.3117551
_cons	78.8292	3.912742	20.15	0.000	71.16037	86.49804

▷ We find that adding another student decreases the average math score by 0.3775 points.

▷ Note that this is the average treatment effect for classes that (1) comply with the cutoff rule $(T = cp)$ and (2) have 40 students $(R = c)$. This is a stricter interpretation than the usual definition of LATE, which is defined for compliers.

Problem 1.5. (*) Use -rdrobust- to estimate the effect of class size on math scores and compare your results.

Solution. Using -rdrobust-, we obtain the following results without controls:

```
Sharp RD estimates using local polynomial regression.

     Cutoff c = -39.5 | Left of c  Right of c            Number of obs =        699
----------------------+----------------------            BW type       =      mserd
        Number of obs |       476         223            Kernel        = Triangular
   Eff. Number of obs |       103          49            VCE method    =         NN
       Order est. (p) |         1           1
       Order bias (q) |         2           2
          BW est. (h) |     4.724       4.724
          BW bias (b) |     7.100       7.100
            rho (h/b) |     0.665       0.665

Outcome: avgmath. Running variable: neg_c_size.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  -2.788     4.1764   -0.6676  0.504   -10.9735     5.39757
            Robust |     -          -     -0.6321  0.527   -13.0605     6.69075
--------------------------------------------------------------------------------
```

and the following results with controls:

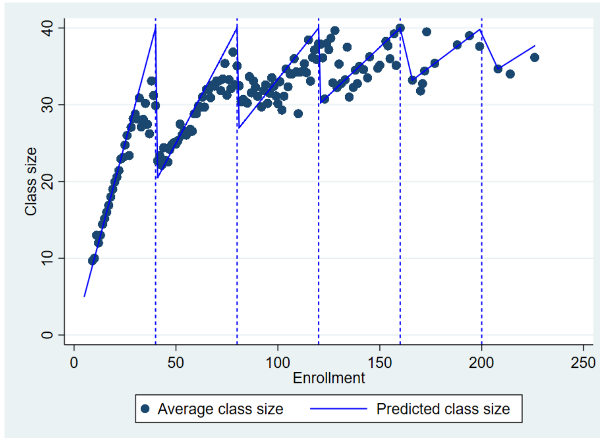
```
       Cutoff c = -40 | Left of c  Right of c            Number of obs =        699
----------------------+----------------------            BW type       =      mserd
        Number of obs |       467         232            Kernel        = Triangular
   Eff. Number of obs |        94          47            VCE method    =         NN
       Order est. (p) |         1           1
       Order bias (q) |         2           2
          BW est. (h) |     4.008       4.008
          BW bias (b) |     7.042       7.042
            rho (h/b) |     0.569       0.569

Outcome: avgmath. Running variable: neg_c_size.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional | -4.7836     2.8489   -1.6791  0.093   -10.3674      .800183
            Robust |     -          -     -1.6340  0.102    -12.622      1.14476
--------------------------------------------------------------------------------
Covariate-adjusted estimates. Additional covariates included: 1
```

▷ Without controls, we find a point estimate of -2.788 absent controls and -4.7836 with controls. This result lines up with the estimates from the OLS and local linear regression in the earlier parts.

Now use the complete sample, and define the following variable predicted class size = enrollment =(int((enrollment − 1)=40) + 1)

Problem 1.6. Plot average class size as a function of enrollment. Add predicted class size to the plot.

Solution. We obtain the following graph:



▷ The dots represent the average class size, and the line represents the predicted class size.

Problem 1.7. Estimate the effect of class size on math scores using IV.

Solution. We obtain the following estimate:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	.2984802	.0469045	6.36	0.000	.2065489	.3904114
_cons	58.38217	1.421793	41.06	0.000	55.59551	61.16884

Instrumented: classize
Instruments: p_classize

▷ We obtain a point estimate of 0.2985, which corresponds to average treatment effect for compliers with class size equal to 40 students.

Problem 1.8. If the RDD is valid, then the coefficient of interest should not change significantly if we include or exclude covariates. Check whether this is the case.

Solution. We obtain the following estimate after including the covariates:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	-.2229847	.0773261	-2.88	0.004	-.3745411	-.0714284
c_size	.042857	.0092301	4.64	0.000	.0247665	.0609476
tipuach	-.3400931	.0167453	-20.31	0.000	-.3729133	-.3072729
_cons	75.46579	1.909501	39.52	0.000	71.72324	79.20834

▷ Now our point estimate is -0.2230 instead of 0.2985 from the previous question.

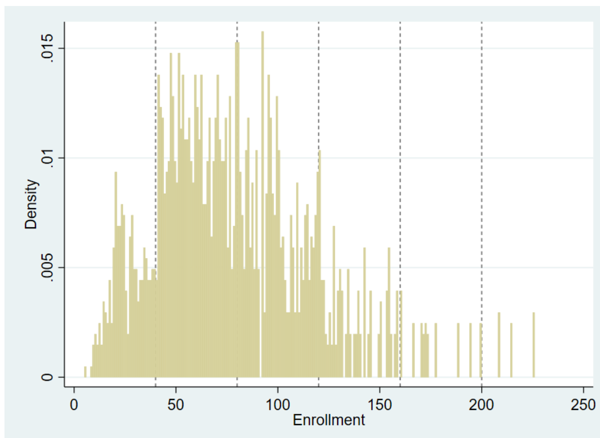
▷ Since the coefficient of interest flips its sign, we may be concerned that the RD design may not be valid.

▷ This could have selection on observables, which raises a concern about potentialselection on unobservables.This could potentially make the regression discontinuity design invalidated.

Explore the validity of the design and the robustness of the results above using the following checks:

Problem 1.9. Manipulation: Plot the distribution of the assignment variable.

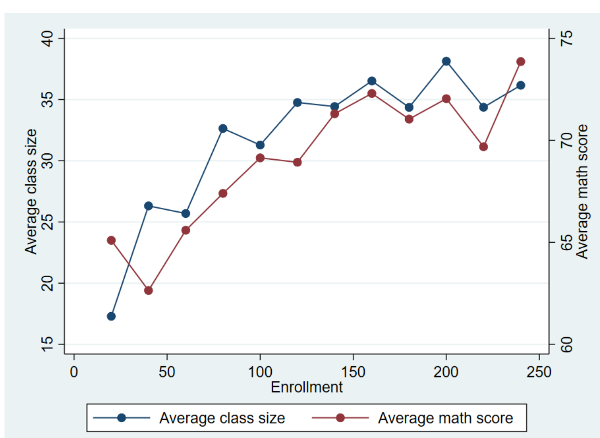
Solution. We obtain the following graph:



▷ If manipulation was the case, we would see sharp jumps at the cutoff. We see a discrete increase at 40.

Problem 1.10. Misspecification 1: Present a graph using binned local averages of class size and math score against enrollment. Use bins of width 20 and make sure that the bins do not cover the discontinuities. Can you see the discontinuity in class size and math scores?

Solution. We obtain the following graph:

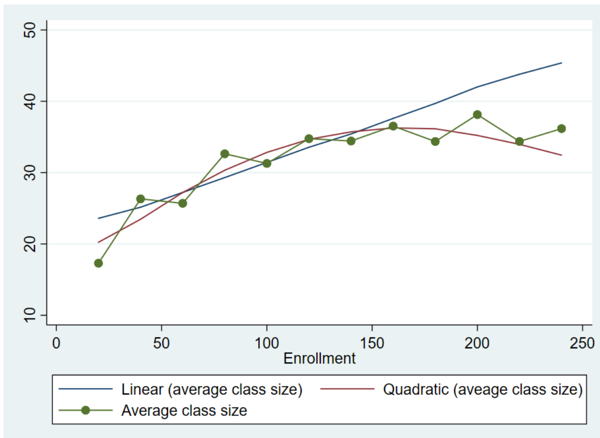


▷ For the average class size, there is a sharp discontinuity at enrollment levels that are multiples of 40.

▷ For the average math score, the discontinuity is less stark.

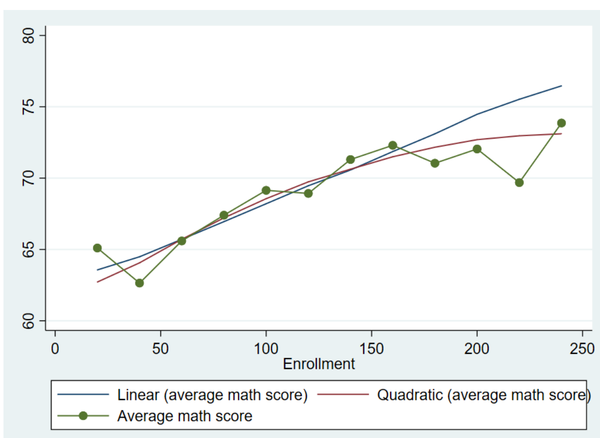
Problem 1.11. Misspecification 2: Superimpose a linear and quadratic trend on the previous graph. Does the polynomial approximation capture the non-linearities well?

Solution. First, we impose trends on the binned local averages of class size against enrollment:



▷ The quadratic trend captures the non-linearities well, while the linear trend does poorly.

Next, we impose trends on the binned local averages of math score against enrollment:



▷ The quadratic trend captures the non-linearities well, while the linear trend does poorly.

Problem 1.12. Misspecification 3: Explore the sensitivity of the results in 7) to 1) bandwidths (restrict the estimation sample to intervals around the discontinuities), and 2) how you control for enrollment.

Solution. First, I try explore the sensitivity of the results respect to the bandwidth absent controls. First for a bandwidth of size 3:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	.5273181	.1164357	4.53	0.000	.2991082	.7555279
_cons	50.83888	3.740506	13.59	0.000	43.50763	58.17014

and bandwith of size 5:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	.3777285	.1001675	3.77	0.000	.1814038	.5740532
_cons	55.78987	3.167504	17.61	0.000	49.58167	61.99806

and bandwidth of size 10:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	.2634784	.0773493	3.41	0.001	.1118765	.4150803
_cons	59.36188	2.433421	24.39	0.000	54.59246	64.1313

We find that the bandwidth choice does affect the estimate.

Second, I control for enrollment in different types. I obtain the following results when I include only a linear term:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	-.2229847	.0773261	-2.88	0.004	-.3745411	-.0714284
tipuach	-.3400931	.0167453	-20.31	0.000	-.3729133	-.3072729
c_size	.042857	.0092301	4.64	0.000	.0247665	.0609476
_cons	75.46579	1.909501	39.52	0.000	71.72324	79.20834

and when I include both a linear and quadratic term:

avgmath	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	-.2767483	.0960186	-2.88	0.004	-.4649413	-.0885553
tipuach	-.3401139	.0168139	-20.23	0.000	-.3730686	-.3071593
c_size	.0786592	.029578	2.66	0.008	.0206873	.1366311
c_size2	-.0001627	.0001188	-1.37	0.171	-.0003955	.0000701
_cons	75.52143	1.928044	39.17	0.000	71.74253	79.30032

I find that adding a quadratic term does not alter the estimate that much.

Problem 1.13. Placebo check: Conduct the RD analysis where your outcome is percentage disadvantaged pupils.

Solution. We run the regression and obtain the following results:

tipuach	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
classize	-.8506549	.0605004	-14.06	0.000	-.9692334	-.7320763
_cons	39.57183	1.833904	21.58	0.000	35.97745	43.16622

We find a significant negative effect of class size on percentage disadvantaged pupils. This brings concerns about potential endogeneity when we are controlling for percentage disadvantaged pupils.