Empirical Methods in Applied Economics
Lecture Notes

Jörn-Steffen Pischke
LSE

October 2007

1 Getting a Little Jumpy: Regression Discontinuity Designs

1.1 The Sharp RD Design

The basic idea of the regression discontinuity (RD) design is extremely simple.[1] It is a straightforward application of the conditional independence assumption $E[\mathrm{Y}_{0i}|\mathrm{X}_i,\mathrm{D}_i] = E[\mathrm{Y}_{0i}|\mathrm{X}_i]$. The key to the RD design is that we have a deep understanding of the mechanism which underlies the assignment of treatment D_i. In this case, there is a single confounder X_i, and in the sharp RD design this variable fully determines the treatment. In particular, the relationship between X_i and the treatment is such that there is a point X_0 and

$$\mathrm{D}_i = \begin{cases} 1 & \text{if } \mathrm{X}_i \geq \mathrm{X}_0 \\ 0 & \text{if } \mathrm{X}_i < \mathrm{X}_0 \end{cases} .$$

Hence, there is a cutoff point for X, and for individuals with X_i above the cutoff point, treatment takes place, while for individuals with X_i below the cutoff point there is no treatment. Figure 1 is a hypothetical illustration where all observations with $\mathrm{X}_i \geq 0.5$ receive the treatment.

Elections are a good example of a situation which creates a regression discontinuity design. In this case, D_i denotes the candidate who is elected, X_i is the vote share for the candidate, and Y_i is an outcome influenced by the elected candidate. A candidate wins the election by receiving more than

[1] The RD design is due to Thistlewaite and Campbell (1960). A special issue of the *Journal of Econometrics* (2007) contains numerous instructive examples. The introduction to this issue by Imbens and Lemieux (2007) provides an excellent survey of the methodology for practitioners.

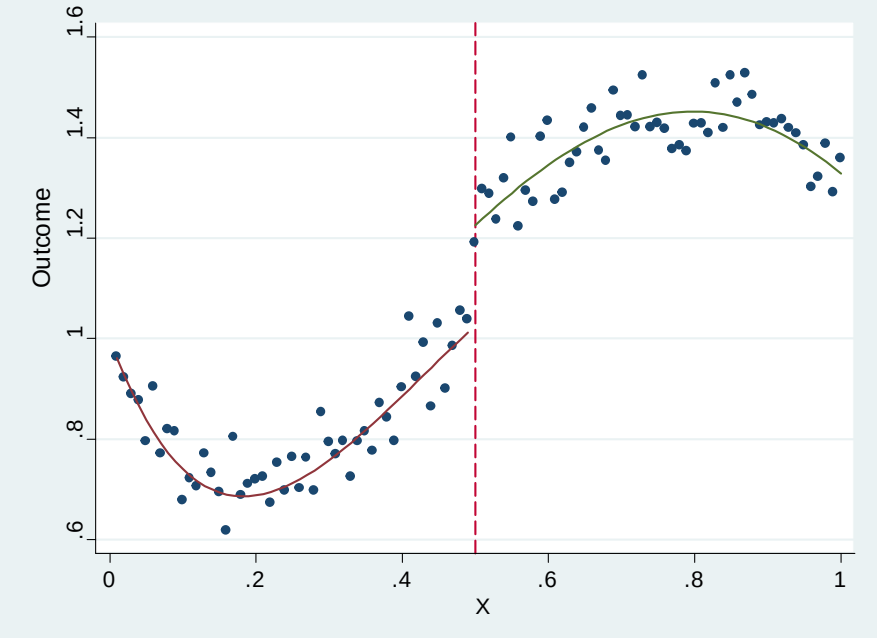


Figure 1: The Sharp Regression Discontinuity Design

50 percent of the vote. The counterfactual outcome, on the other hand, changes continuously with the vote share, so it should be very similar for a vote share of 49 percent or 51 percent, although a different candidate is elected.

Figure 1 suggests an obvious way to identify the causal effect in this case. $E[\mathrm{Y}_{0i}|\mathrm{X}_i]$ will be some function of X_i, so let's write

$$\begin{aligned} \mathrm{Y}_{0i} &= f(\mathrm{X}_i) + \varepsilon_i \\ \mathrm{Y}_{1i} &= \mathrm{Y}_{0i} + \beta \\ \mathrm{Y}_i &= f(\mathrm{X}_i) + \beta \mathrm{D}_i + \varepsilon_i \qquad (1) \\ &= f(\mathrm{X}_i) + \beta 1(\mathrm{X}_i \geq \mathrm{X}_0) + \varepsilon_i. \qquad (2) \end{aligned}$$

The function $f(\cdot)$ must be continuous at X_0 to avoid collinearity in the model, and this is the basic assumption of the RD design. In practice, we will have to assume some flexible functional form for $f(\cdot)$, for example a polynomial. We can then simply run the regression (2), and this is typically the starting point for an RD analysis (together with plotting the data as in figure 1).

If we have chosen $f(\cdot)$ flexibly enough then this regression is identifed from the observations around the point of discontinuity, X_0. Look at figure 1 again, and let δ be a small postitive number. Then

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{X}_0-\delta<\mathrm{X}_i<\mathrm{X}_0\right] &\simeq E[\mathrm{Y}_{0i}|\mathrm{X}_0] \\ E\left[\mathrm{Y}_i|\mathrm{X}_0<\mathrm{X}_i<\mathrm{X}_0+\delta\right] &\simeq E[\mathrm{Y}_{1i}|\mathrm{X}_0] \end{aligned}$$

and hence

$$\lim_{\delta\to 0} E\left[\mathrm{Y}_i|\mathrm{X}_0<\mathrm{X}_i<\mathrm{X}_0+\delta\right]-E\left[\mathrm{Y}_i|\mathrm{X}_0-\delta<\mathrm{X}_i<\mathrm{X}_0\right]=E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{X}_0].$$

This says that if we compare the mean outcome in a small neighborhood to the left and to the right of the point of discontinuity this gives us an estimate of the treatment effect at the point of the discontinuity. However, if the treatment effects are different at different points of X we will always only be able to get a local treatment effect at the point X_0. The RD estimate is by its nature a local estimate.

In parctice, using the mean of Y_i will lead to a biased estimate whenever the slope of $E[\mathrm{Y}_{0i}|\mathrm{X}_i]$ is non-zero because we are interested in the estimate at the boundary, and we only use data from one side of that boundary. Nevertheless, the insight that the data points around the cutoff value X_0 provide the essential identfying information is an important one, and it suggests to restrict the analysis to a discontinuity sample around X_0. In order to avoid the boundary problem it is advisable to still use regression even within this smaller sample (Hahn, Tood, and van der Klaauw, 2001, and Porter, 2003). A lower order polynomial or a simple linear regression is now typically sufficient.

The regression (2) only identifies the treatment effect if $f(\mathrm{X}_i)$ is not just continuous but also smooth around the point X_0, and if the treatment effect is constant. While smoothness is often a reasonable assumption, constant treatment effects certainly often is not. If $E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{X}]$ depends on X, then the slope of $E[\mathrm{Y}_i|\mathrm{X}_i]=f(\mathrm{X}_i)+E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{X}_i]\mathrm{D}_i$ may differ to the right and to the left of X_0. So it makes sense to use the subsamples $\{\mathrm{X}_0<\mathrm{X}_i<\mathrm{X}_0+\delta\}$ and $\{\mathrm{X}_0-\delta<\mathrm{X}_i<\mathrm{X}_0\}$ separately. The difference of the two estimates $\widehat{f}_l(\mathrm{X}_0)$ (from the sample to the left of X_0) and $\widehat{f}_r(\mathrm{X}_0)$ (from the sample to the right of X_0) gives us our estimate of the treatment effect. We can also pool these two regressions into one and run

$$\mathrm{Y}_i=f_l(\mathrm{X}_i)1(\mathrm{X}_i<\mathrm{X}_0)+f_r(\mathrm{X}_i)1(\mathrm{X}_i\geq\mathrm{X}_0)+\beta 1(\mathrm{X}_i\geq\mathrm{X}_0)+\varepsilon_i \qquad (3)$$

In this specification, it is important to specify the two functions $f_l(\cdot)$ and $f_r(\cdot)$ so that $f_l(\mathrm{X}_0) = f_r(\mathrm{X}_0)$ holds, otherwise β does not capture the jump at the discontinuity correctly. In this case, the formulation (3) will still estimate $\beta = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{X}_0]$ as the local effect. Any deviation of the treatment effect from β further away from the cutoff point is modeled by $f_r(\mathrm{X}_i)$.

Imbens and Lemieux (2007) suggest that using a linear function for $f(\mathrm{X}_i)$ in a small enough sample around X_0 should suffice. In this case we run the regression

$$\mathrm{Y}_i = \alpha + \gamma_1 1(\mathrm{X}_i < \mathrm{X}_0)(\mathrm{X}_i - \mathrm{X}_0) + \gamma_2 1(\mathrm{X}_i \geq \mathrm{X}_0)(\mathrm{X}_i - \mathrm{X}_0) + \beta \mathrm{D}_i + \varepsilon_i. \quad (4)$$

The parametrization using $(\mathrm{X}_i - \mathrm{X}_0)$ forces the constant α to capture the location of the mean of Y_i just to the left of X_0, and $f_l(\mathrm{X}_i) = \gamma_1 1(\mathrm{X}_i < \mathrm{X}_0)(\mathrm{X}_i - \mathrm{X}_0)$ and $f_r(\mathrm{X}_i) = \gamma_2 1(\mathrm{X}_i \geq \mathrm{X}_0)(\mathrm{X}_i - \mathrm{X}_0)$ both meet at that point. β therefore estimates the treatment effect at X_0. The difference $\gamma_2 - \gamma_1$ is part of the control, although this difference may well capture the non-constancy of the treatment effect away from X_0. Conventional standard errors for this regression are appropriate.

A final question is how big a sample around the discontinuity to use. There is the typical tradeoff between bias and precision. Only the observations close to X_0 really carry information on what's going on at that point. Using more observations increases the precision of the estimate but at the cost of extrapolating from (the less informative) observations further away from the discontinuity. Imbens and Lemieux (2007) suggest a simple method for choosing an optimal sample size. In practice, we feel that it suffices to use rules of thumb to choose the size of the discontinuity sample. More important than getting this sample size just right is doing some experimentation with smaller and larger samples, and different polynomials for $f(\cdot)$ in the larger samples.

An example of the sharp regression discontinuity design using the cutoff rule inherent in elections is the study by Lee (2007) of the advantage of incumbency on re-election probabilities. The question of interest is whether a politician holding a particular office is more likely to win a future election than a challenger not holding the same political office. Incumbent politicians may have resources at their disposal which help them in gaining re-election compared to challengers. On the other hand, incumbent politicians have been elected in the past, and hence may also simply be prefered by voters. The idea of the Lee study is that voters' preferences will be summarized by a politician's vote share in the past election. Hence, comparing politicians who just won a past election (and hence became incumbent politicians for a

future election) to those who just lost a past election (and hence do not hold the same political office) can give an estimate of the incumbancy effect. This is a regression discontinuity setup: the vote share in the past election is the variable X_i, which captures the confounding variable on voters preferences. The treatment is victory in the past election (and hence incumbancy) and the outcome is the probability of winning a future election.

The key feature of the RD design is that only a single variable determines the selection rule, and that variable is known and available to us as researchers. Then it suffices to control for this single confounder X_i. For example, a politician may be more attractive to voters because he is good looking. This will increase the probability of winning the election in $t+1$ but it also affected the probability of victory in election t. Victory in t, of course, was also affected by other factors, like the strength of the opponent in that election, and all of this gets lumped into the vote share X_i. So isn't there more information in other variables, like physical attractiveness of the candidate, which would help forcast the repeat election in $t+1$? There is, of course. But the key insight is that none of this information is correlated with incumbancy *in expected value*, i.e. across a large number of elections, around the point of discontinuity.

Figure 2 from Lee (2007) illustrates the results of this exercise for elections to the US House of Representatives. The top figure plots the probability of winning the election in $t+1$ against the vote share in election t. Future election probabilities are a function of the past vote share, but there is a large discrete jump at the point where the candidate wins the election in the past. The figure shows that incumbancy raises the re-election probability by about 35 percentage points. Figure 2b checks the identification strategy by looking at the number of past election victories before election t as the outcome. This should not be affected by winning the election t, and this is indeed born out by the data. Checking the RD design for jumps in variables which should not be affected by the treatment always raises the confidence in the design. An additional useful check is to look at an estimate of the density of X_i around the point of discontinuity, particularly if there is a worry that individuals might manipulate this variable in response to the threshold. This can typically be done by showing a simple histogram.

1.2 Fuzzy RD is IV

The design described so far is called the sharp regression discontinuity design because D_i changes discretely at the point X_0. This is, however, not necessary for the regression discontinuity design. It is enough that the prob-

ability of treatment assignment changes discretely at X_0, i.e. $\lim_{\delta \to 0} P(D_i = 1|X_0 - \delta) < \lim_{\delta \to 0} P(D_i = 1|X_0 + \delta)$. Now, there will be both treated or untreated observations on either side of the point of discontinuity. Hence,

$$\begin{aligned} & E\left[Y_i|X_0 < X_i < X_0 + \delta\right] - E\left[Y_i|X_0 - \delta < X_i < X_0\right] \\ \simeq \quad & \beta\left(P\left[D_i|X_0 < X_i < X_0 + \delta\right] - P\left[D_i|X_0 - \delta < X_i < X_0\right]\right) \end{aligned}$$

$$\Rightarrow \frac{E\left[Y_i|X_0 < X_i < X_0 + \delta\right] - E\left[Y_i|X_0 - \delta < X_i < X_0\right]}{E\left[D_i|X_0 < X_i < X_0 + \delta\right] - E\left[D_i|X_0 - \delta < X_i < X_0\right]} \simeq \beta.$$

It is easy to see that this is the Wald estimator, with the indicator $1(X_i \geq X_0)$ as the instrumental variable for the treatment assignment D_i.[2] Similarly, we can estimate equation (1) or (4) by instrumental variables with $1(X_i \geq X_0)$ as the instrument for D_i. (1) and (2) are now no longer identical. (1) is the structural equation, while (2) is the reduced form.

An example of a fuzzy regression discontinuity design with a continuous treatment variable is the study of the effect of class size on student performance by Angrist and Lavy (1999). There study uses the fact that class size in Israeli schools is capped at 40. Once enrollment reaches 41, two classes are formed, three classes at 81 and so on (this rule goes back to the medieval talmudic scholar Maimonides and is hence referred to by the authors as "Maimonides rule"). This example is more general than what we have discussed so far in two more repects. First, class size is a function of enrollment, with discontinuties at multiples of 40. So there are discontinuities at all enrollment multiples of 40, and these are easily exploited together. Second, class size, the treatment variable of interest is continuous rather than binary. Since the RD design only provides local estimates around the discontinuity X_0, with a single discontinuity, say at an enrollment of 80, we would only learn about the difference in class sizes of 40 and 27. With the multiple discontinuities we can, at least in principle, learn something about various points in the class size-performance relationship.

The problem is again that enrollment may correlate with student performance indepently of its effect on class size. For example, bigger schools are more likely to be in urban areas with a better intake of students. Bigger schools may also offer more economies of scale, and hence make better use

[2] The fuzzy RD design is due to Trochim (1984). Angrist and Lavy (1999) and van der Klaauw (2002) first related this design to IV methods. Hahn, Todd, and van der Klaauw (2001) provide a theoretical treatment.

of resources. It is therefore important to be able to control for enrollment directly. Figure 1 from Angrist and Lavy (1999) plots actual class sizes against enrollment, and also shows predicted class sizes from Maimonides' rule. The class size patterns roughly follow Maimonides' rule, although the fit is not exact because sometimes schools split classes before reaching the maximum size of 40. However, there are clearly visible declines in class sizes at enrollment levels of 40, 80, and 120. Hence, the regression discontinuity design here is fuzzy, rather than sharp, and the correct empirical implementation is to instrument class size with Maimonides' rule.

Figure 2 plots predicted class size against average reading scores by enrollment. The plots show an inverse saw tooth pattern, suggesting that smaller classes may be good for performance. It also demonstrates that students in larger schools do better on average. Tables 2 and 4 shows some estimates of the class size effect on math performance of 5th graders. The OLS results demonstrate that there is a positive correlation between class size and test scores in the raw data. This correlation vanishes when the fraction disadvantaged students is controlled for. The IV results exploit the regression discontinuities created by Maimonides' rule. The table displays various specifications with no control for enrollment, and with linear, and quadratic controls for enrollment, as well as estimates in subsamples around the discontinuity points. Controlling for enrollment is important as can be seen by the comparison of columns (4) and (5) of Table 2. The form of the control matters less. On the other hand, the discontinuity samples gives sometimes larger effects (in absolute values) than the full sample, but the standard errors are fairly large as well.

2 References

Angrist, Joshua, and Victor Lavy, Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement, Quarterly Journal of Economics Volume (Year): 114 (1999), Issue (Month): 2 (May), Pages: 533-575

David S. Lee, Randomized experiments from non-random selection in U.S. House elections, Journal of Econometrics, 2007

Hahn, Jinyong, Petra Todd, and Wilbur van der Klaauw (2001): Identification and Estimation of Treatment Effects with a Regression-Discontinuty Design, Econometrica 69, 201-209.

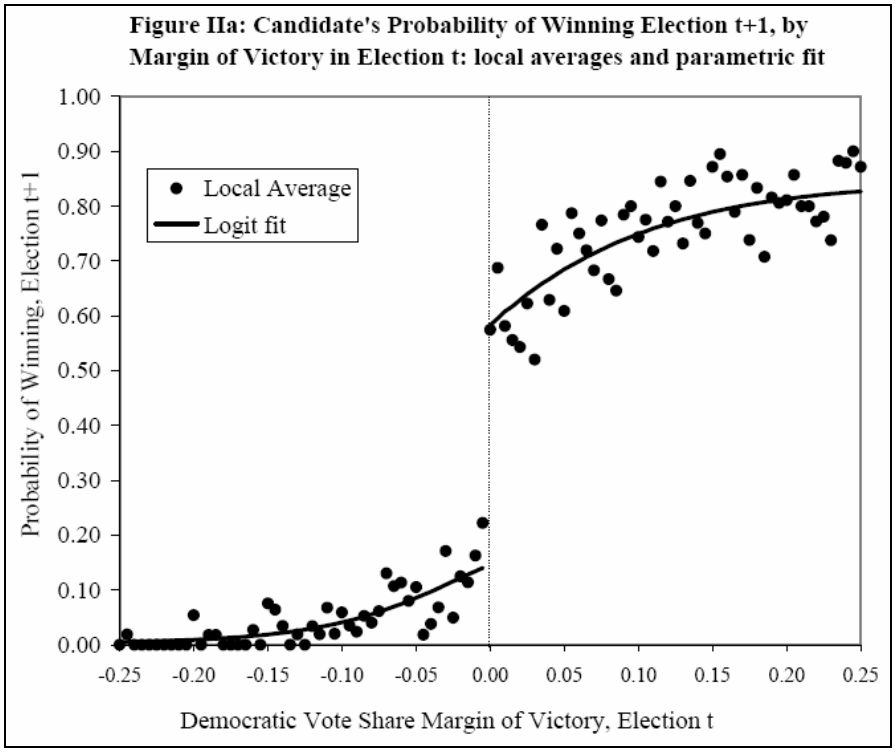
Imbens, Guido, and Thomas Lemieux (2007) Regression Discontinuity Designs: A Guide to Practice, Journal of Econometrics, 2007

Thistlewaite, D, and D. Campbell (1960) Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment, Journal of Educational Psychology, 51, 309-317.

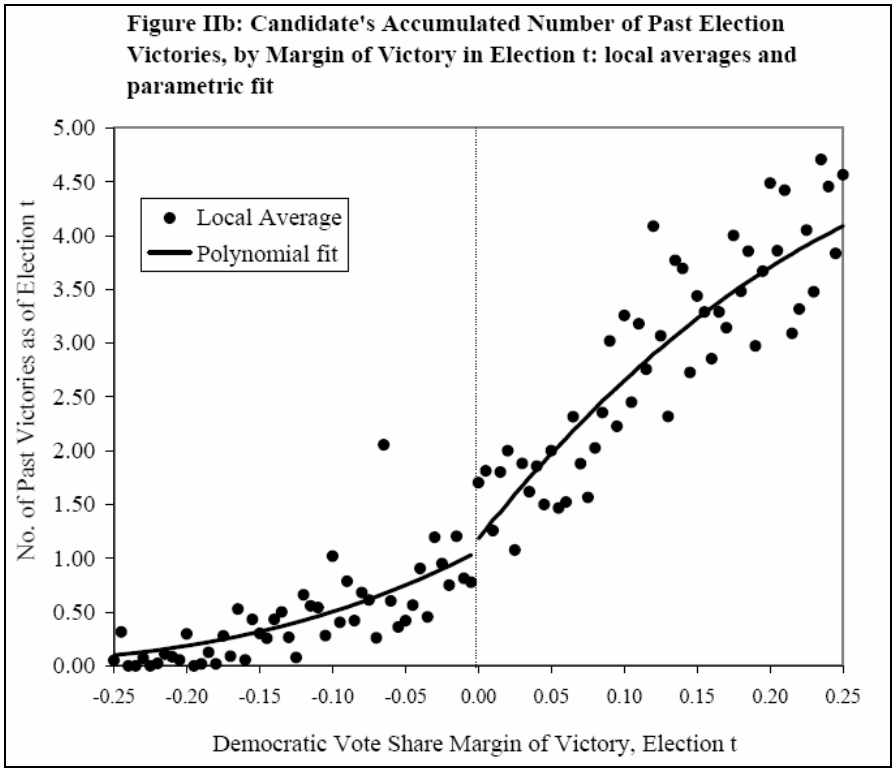
Trochim. W (1984) Research Designs for Program Evaluation. The Regression Discontinuity Design. Beverly Hills, CA: Sage Publications

Wilbert van der Klaauw (2002) Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach, International Economic Review, Vol 43(4), November 2002.

Lee 2005: Figure 2a



Lee 2005: Figure 2b



Angrist and Lavy 1999: Figure 1

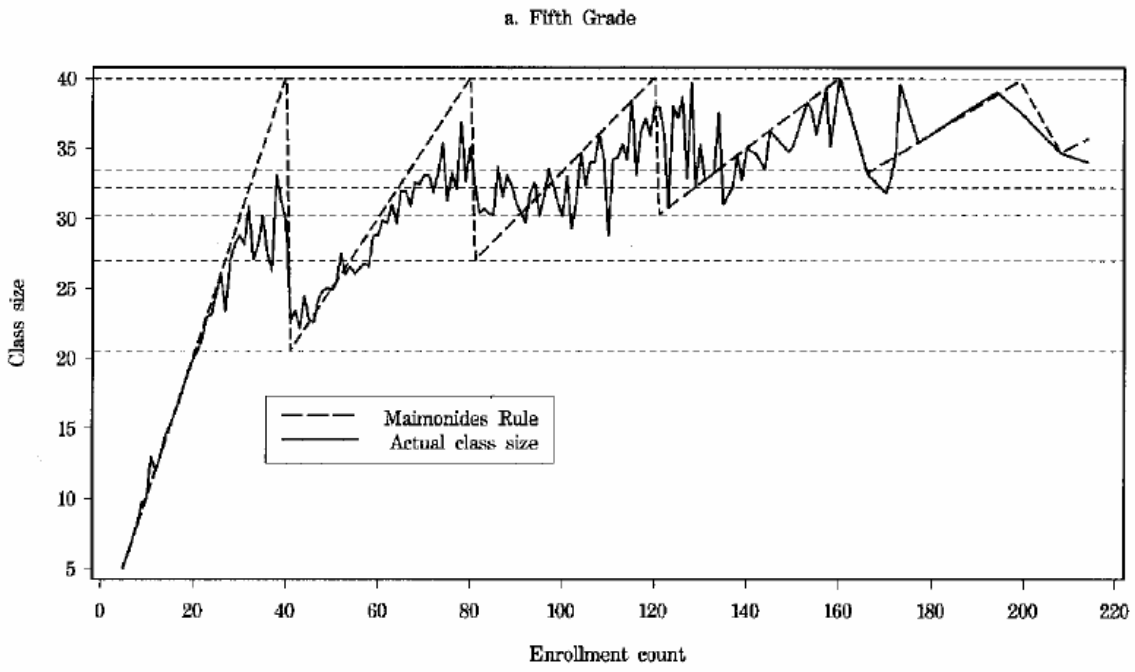


FIGURE I
Class Size in 1991 by Initial Enrollment Count, Actual Average Size and as Predicted by Maimonides' Rule

Angrist and Lavy 1999: Figure 2

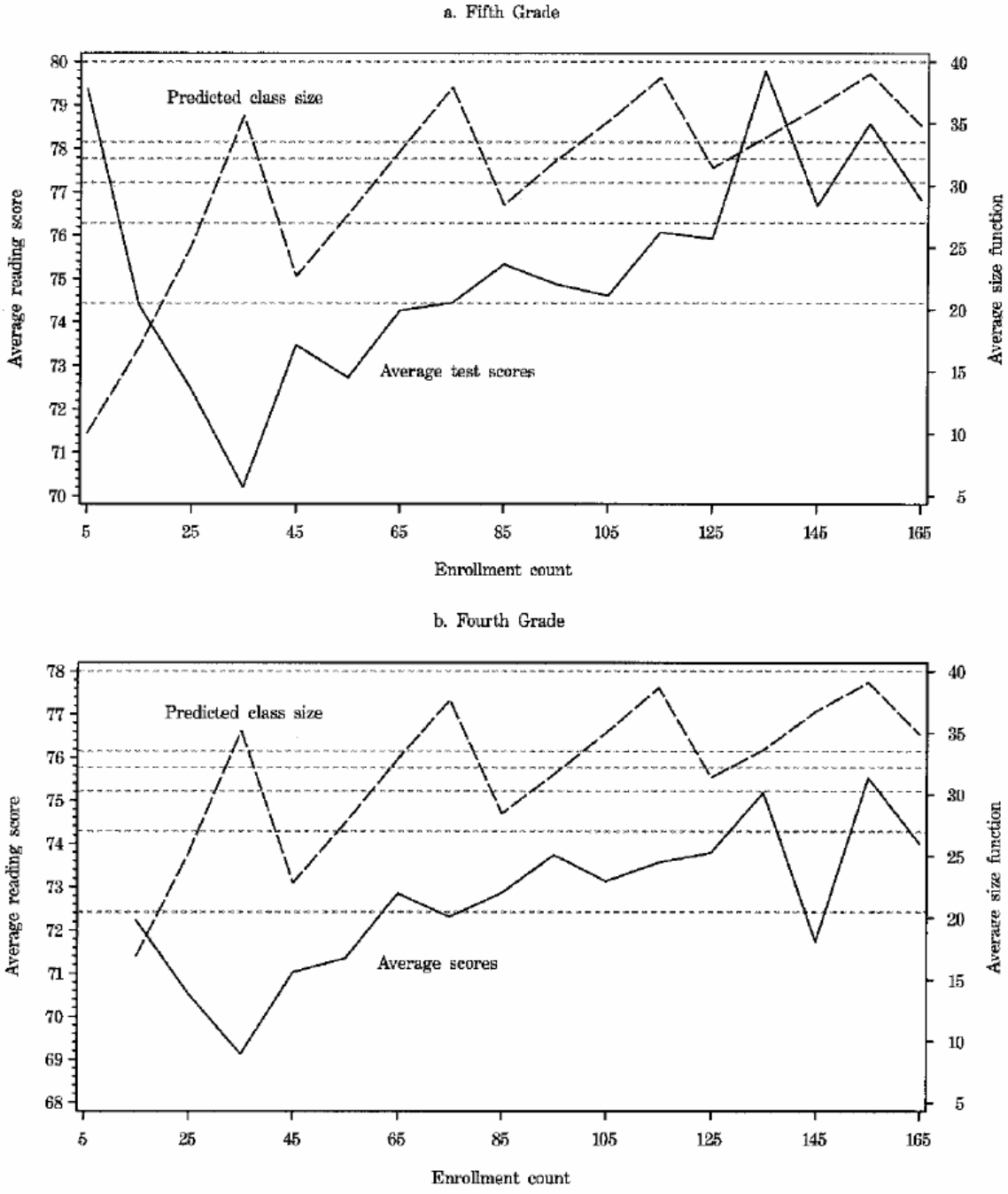


FIGURE II
Average Reading Scores by Enrollment Count, and the Corresponding Average Class Size Predicted by Maimonides' Rule

Angrist and Lavy 1999: Table 2

TABLE II
OLS ESTIMATES FOR 1991

	5th Grade						4th Grade					
	Reading comprehension			Math			Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Mean score		74.3			67.3			72.5			69.9	
(s.d.)		(8.1)			(9.9)			(8.0)			(8.8)	
Regressors												
Class size	.221	−.031	−.025	.322	.076	.019	0.141	−.053	−.040	.221	.055	.009
	(.031)	(.026)	(.031)	(.039)	(.036)	(.044)	(.033)	(.028)	(.033)	(.036)	(.033)	(.039)
Percent disadvantaged		−.350	−.351		−.340	−.332		−.339	−.341		−.289	−.281
		(.012)	(.013)		(.018)	(.018)		(.013)	(.014)		(.016)	(.016)
Enrollment			−.002			.017			−.004			.014
			(.006)			(.009)			(.007)			(.008)
Root MSE	7.54	6.10	6.10	9.36	8.32	8.30	7.94	6.65	6.65	8.66	7.82	7.81
R^2	.036	.369	.369	.048	.249	.252	.013	.309	.309	.025	.204	.207
N		2,019			2,018			2,049			2,049	

The unit of observation is the average score in the class. Standard errors are reported in parentheses. Standard errors were corrected for within-school correlation between classes.

Angrist and Lavy 1999: Table 4

TABLE IV
2SLS ESTIMATES FOR 1991 (FIFTH GRADERS)

	Reading comprehension						Math					
	Full sample				+/− 5 Discontinuity sample		Full sample				+/− 5 Discontinuity sample	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Mean score		74.4			74.5			67.3			67.0	
(s.d.)		(7.7)			(8.2)			(9.6)			(10.2)	
Regressors												
Class size	−.158	−.275	−.260	−.186	−.410	−.582	−.013	−.230	−.261	−.202	−.185	−.443
	(.040)	(.066)	(.081)	(.104)	(.113)	(.181)	(.056)	(.092)	(.113)	(.131)	(.151)	(.236)
Percent disadvantaged	−.372	−.369	−.369		−.477	−.461	−.355	−.350	−.350		−.459	−.435
	(.014)	(.014)	(.013)		(.037)	(.037)	(.019)	(.019)	(.019)		(.049)	(.049)
Enrollment		.022	.012			.053		.041	.062			.079
		(.009)	(.026)			(.028)		(.012)	(.037)			(.036)
Enrollment squared/100			.005						−.010			
			(.011)						(.016)			
Piecewise linear trend				.136						.193		
				(.032)						(.040)		
Root MSE	6.15	6.23	6.22	7.71	6.79	7.15	8.34	8.40	8.42	9.49	8.79	9.10
N		2019		1961	471			2018		1960	471	

The unit of observation is the average score in the class. Standard errors are reported in parentheses. Standard errors were corrected for within-school correlation between classes. All estimates use f_{ic} as an instrument for class size.