Evidence on the deleterious impact of sustained use of polynomial regression on causal inference*

Andrew Gelman†

2 Mar 2014

Abstract

It is standard in regression discontinuity analysis to control for third- or fifth-degree polynomials of the assignment variable. Such models can overfit, leading to causal inferences that are substantively implausible in which variation is somewhat arbitrarily attributed to high-degree polynomial or the discontinuity. We discuss in the context of a recent study of the effect on air pollution and life expectancy coming from a coal-heating policy in China. In successful regression discontinuity analyses, the underlying relation between the assignment variable and the outcome tends to be strong, monotonic, and with a clear substantive justification. Without such a clear underlying pattern, the validity estimated coefficient from the discontinuity is much less clear. We recommend that (a) researchers recognize these problems which can be exacerbated by controlling for higher-order polynomials; and (b) that journals recognize that quantitative analyses of policy issues are often inconclusive and relax the implicit rule under which statistical significance is a condition for publication.

Keywords: identification, policy analysis, polynomial regression, regression discontinuity, uncertainty

1. Regression discontinuity analysis

Causal inference is central to science, and identification strategies are central to causal inference in aspects of social and natural sciences where large-scale experimentation is not possible. Regression discontinuity (RD) methods are a longstanding and recently popular way to get identification in settings with natural experiments. But implementations of regression discontinuity inference can have serious problems in practice, especially with the often-standard approach of controlling for high-degree polynomials of the underlying continuous predictor.

We demonstrate with a recent well-publicized example in public health where a high-degree polynomial control in a RD analyses led to implausible conclusions. In addition to implying that researchers should show much more caution with such models, this experience suggests a rethinking of conventional ideas of robustness to model specification.

Regression discontinuity analysis has analysis enjoyed a renaissance, especially in economics; Lee and Lemieux (2010) provide an influential review. In their words, RD is "a way of estimating treatment effects in a nonexperimental setting where treatment is determined by whether an observed 'assignment' variable (also referred to in the literature as the 'forcing' variable or the 'running' variable) exceeds a known cutoff point." To the extent that the assignment depends (perhaps stochastically) only on this rule, and to the extent that there are no systematic pre-treatment differences between the items below and above the cutoff, the RD design can be interpreted as a quasi-experiment and the resulting inferences can be interpreted causally.

Lemieux and Lee continue: "RD designs require seemingly mild assumptions compared to those needed for other nonexperimental approaches. . . . the belief that the RD design is not 'just another' evaluation strategy, and that causal inferences from RD designs are potentially more credible than those from typical 'natural experiment' strategies . . . This notion has a theoretical justification . . . "

*We thank Jennifer Hill for helpful comments.

†Department of Statistics, Columbia University, New York.

One way to see the appeal of RD is to consider the threats to validity that arise with five other methods used for causal inference in observational studies: simple regression, matching, selection modeling, difference in differences, and instrumental variables. These competitors to RD all have serious limitations: regression with many predictors becomes model dependent (using the least squares approaches that are traditional in econometrics, it is difficult to control for large numbers of predictors, while nonparametric approaches such as Bart (Hill, 2013) have not yet gained wide acceptance); matching, like linear or nonlinear regression adjustment, leans on the assumption that treatment assignment is ignorable conditional on the variables used to match; selection modeling is sensitive to untestable distributional assumptions; difference in differences requires an additive model that is not generally plausible; and instrumental variables, of course, only work when there happens to be a good instrument related to the causal question of interest.

For all these reasons, in practice causal analyses often seem to flow from identification opportunities to inferences of interest (Gelman, 2009), a view that contrasts with the usual textbook presentation in which the research question comes first and then the analyst finds an identification strategy to attack the problem at hand.

Many of the challenges of applying an identification strategy arise in the data analysis. Sample sizes can be small (especially in areas such as political science or economics where one cannot simply augment a dataset by instigating a few more wars, scandals, or recessions), and theoretical results of unbiasedness do not always help much, first because low bias has no practical meaning in the presence of high variance, and second because datasets are typically constructed by pooling over different subpopulations or different time periods or different sorts of cases, so that any claims of unbiased estimates typically apply only to aggregates that are not directly relevant to the ultimate questions of interest.

For this reasons, the Lee and Lemieux paper is welcome in that it continually returns to practical issues of estimation. Particularly relevant for the purposes of our discussion here are two of their recommendations:

- "From an applied perspective, a simple way of relaxing the linearity assumption is to include polynomial functions of X in the regression model. . . . it is advisable to try and report a number of specifications to see to what extent the results are sensitive to the order of the polynomial."

- "Graphical presentation of an RD design is helpful and informative but the visual presentation should not be tilted toward either finding an effect or finding no effect."

Both these pieces of advice seem reasonable (although, in the first case, we would prefer a spline or Gaussian process or some other such smooth model, as indeed has been suggested by some econometricians). The challenge is what to do *after* following this advice.

2. Example: A claim that coal heating is reducing lifespan by 5 years for half a billion people

We discuss in the context of a paper by Chen et al. (2013) that received a lot of attention with the following claim:

> This paper's findings suggest that an arbitrary Chinese policy that greatly increases total suspended particulates (TSPs) air pollution is causing the 500 million residents of Northern China to lose more than 2.5 billion life years of life expectancy. The quasi-experimental empirical approach is based on China's Huai River policy, which provided

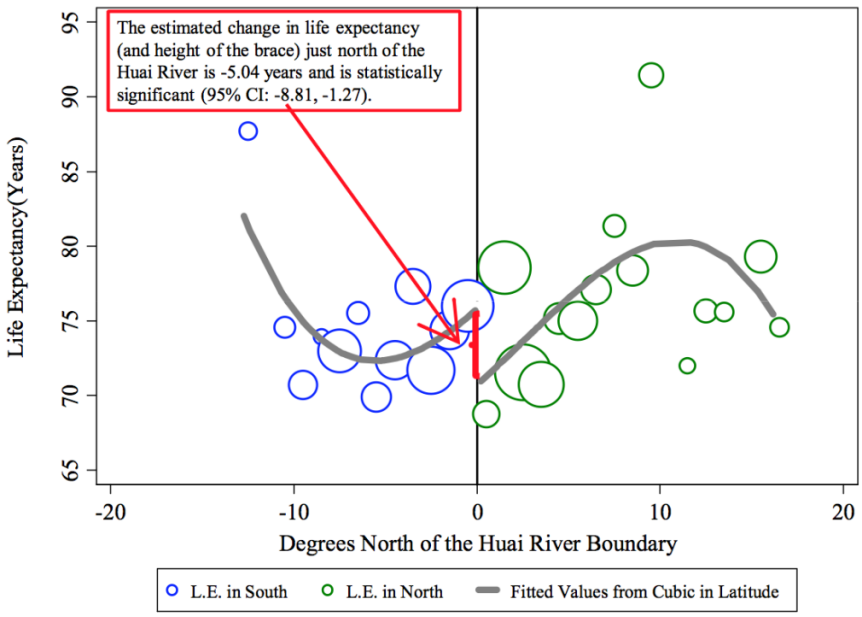

Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Figure 1: *Key graph from Chen et al. (2013) showing their regression discontinuity analysis. Each circle represents a province in China.*

> free winter heating via the provision of coal for boilers in cities north of the Huai River but denied heat to the south. Using a regression discontinuity design based on distance from the Huai River, we find that ambient concentrations of TSPs are about 184 μg/m3 [95% confidence interval (CI): 61, 307] or 55% higher in the north. Further, the results indicate that life expectancies are about 5.5 y (95% CI: 0.8, 10.2) lower in the north owing to an increased incidence of cardiorespiratory mortality.

Before going on, let us just say that, if you buy this result, you should still be interested in it even if the 95% confidence intervals had happened to include zero. There is an unfortunate convention that "p less than .05" results are publishable while "non-significant" results are not. The life expectancy of 500 million people is important, and it's inappropriate to wait on statistical significance to make policy decisions in this area.

Getting to the details, though, we are far less than 97.5% sure that the effects are in the direction that the authors claim (recall that 97.5% is the posterior probability of a positive effect given $p = .05$, under a flat prior). And, for the usual proper-prior Bayesian reasons, we would guess that this "2.5 billion years of life expectancy" is an overestimate: great swathes of the 95% confidence interval represent very large effects that seem a priori unlikely.

We have reproduced the key graph of Chen et al. as Figure 1 here. It is a beautiful graph, showing the model and the data together and following the advice or Lee and Lemieux reported above. One particular admirable aspect of this particular graph is that, just looking at it, you can see how odd the model is. Or, at least, how odd it looks to an outsider. A third-degree polynomial indeed! It looks like that's where the claim of 5 years of life expectancy came from. We remain a little confused still, because the interval is [1.3, 8.1] in the graph and [0.8, 10.2] in the abstract, so

Table S9

Robustness checks of choice of functional form for latitude

	Linear & Controls	Quadratic & Controls	Cubic & Controls	Quartic & Controls	Quintic & Controls
	(1)	(2)	(3)	(4)	(5)
Panel 1: Impact of "North" on the Listed Variable, Ordinary Least Squares					
TSP (100 $\mu g/m^3$)	2.89***	2.63***	1.84***	1.95***	1.52**
	(0.56)	(0.49)	(0.63)	(0.59)	(0.72)

Figure 2: *Part of a table from the supplemental material for Chen et al. (2013). The original caption reads, "Robustness checks of choice of functional form for latitude." Our problem with all these models is that they do not include other predictors and the residual errors are large (see Figure 1), hence the causal estimate based on regression discontinuity is highly sensitive to the assumption that the other factors (represented by a combination of the nonlinearity and the error term in the regression model) are uncorrelated with the discontinuity.*

there must be something else going on, but this seems like the basic story.

Table S.9 in the supplemental material gives the authors' results trying other models. The cubic adjustment gave an estimated effect of 5.5 years with standard error 2.4. A linear adjustment gave an estimate of 1.6 years with standard error 1.7. Figure 2 here shows the relevant part of the table.

Our point here is not to argue that the linear model is correct—the authors in fact supply data-based reasons for preferring the cubic—but rather that the headline claim, and its statistical significance, is highly dependent on a model choice that has no particular *scientific* (as distinguished from data-analytic) basis. Figure 1 above indicates to us that neither the linear nor the cubic nor any other polynomial model is appropriate here; that there are other variables not included in the model that distinguish the circles in the graph. A multilevel model might be a good idea; of course that would increase standard errors in its own way. Or one could try some less model-based approach such as the robust regression discontinuity method of Calonico, Cattaneo, and Titiunik. We prefer modeling because, for us, the model-building step meshes with the goal of increasing substantive understanding. But we recognize that different philosophies can converge to similar results. In any case, we would like to ditch that approach of estimating high-degree polynomials.

3. Discussion

3.1. Publication of speculative findings on particulate pollution and life expectancy

There might well be good reasons for expecting an effect of even more than 5 years of life expectancy from this policy—we are not experts on particulate pollution and health—but, from the above data alone, the claim of 5 years looks artifactual. And, there also seems to be an implication that, in those northern areas with life expectancy around 80, that people would be living to 85 in the absence of the policy (see Sumner, 2013, for a further discussion of this point). But an average lifespan of 85 seems like a strong conclusion to come to, if it's being driven by this data analysis alone. Which is what Chen et al. seem to be doing, in that they're just taking their estimated regression coefficient and considering it as a treatment effect.

We are not saying that particulate matter doesn't kill, or that this topic shouldn't be studied, or that these findings shouldn't be published in a high-profile journal. The accompanying article by Pope and Dockery (2013) considers why Chen et al.'s conclusions might be scientifically plausible.

Rather, what we are suggesting is a two-step: the authors retreat from their strongly model-based claim of statistical significance, and the journal accept that non-statistically-significant findings on important topics are still worth publishing.

The underlying relationship in Figure 1 is not anything close to smooth; the cities differ in their mortality rates in important ways not related to latitude. This in turn points to a crucial problem with the RD analysis: the "natural experiment" can in fact be strongly correlated with important unobserved (and non-random) city-level predictors. And thus, absent a plausible underlying relationship of y on x, RD shares all the difficulties of any observational study. Conditional on having an assignment variable that is strongly predictive of the outcome, we would like to see something monotonic and with a clear substantive justification.

3.2. Plausibility of a regression discontinuity estimate in the context of the model

At a more technical level, we understand the appeal of controlling for high-order polynomials of the assignment variable, following the general principle that it is safest to control for potential confounders. The usual story is that including one more background variable in the regression model reduces bias with the only cost being a possible increase in variance. Researchers are generally more concerned about bias than variance (especially in a setting such as the one discussed here where estimates are more than two standard errors from zero; see Figure 2), hence it can seem like the conservative choice to control for higher-order polynomials, to reduce whatever bias might arise from curvature in the response function near the discontinuity point.

This reasoning in terms of bias, however, does not always work. And, more to the point, problems can be apparent in particular cases. Again, return to Figure 1, which reveals how much of the estimated discontinuity arises from the steep gradient estimated in life expectancy with latitude near the discontinuity. It would seem to be difficult to understand the purported discontinuity without understanding the gradient (and understanding why that gradient is so strong only in a zone near the boundary).

In many other regression discontinuity studies, the underlying predictive effect of the assignment variable is more clear. For example, in the Lee (2004) study of incumbent party and elections, it makes perfect sense that there will be an approximately linear relation between Democratic or Republican shares in one election and the next; and in the Berger and Pope (2011) study of motivation in basketball, the probability of winning the game is unsurprisingly strongly, smoothly, and monotonically predicted by the score differential at halftime. Also in both these cases, the fit from a cubic polynomial is not far from a straight line on the original or logistic scale. Thus, the higher-order polynomial has the effect of slightly modifying and improving the fit of the natural linear model. In contrast, the scale of both the curve and the discontinuity in Figure 1 lack face validity and raise the possibility of large effects, not accounted for in the model, that happen to have strong statistical dependence with the assignment variable.

3.3. Skepticism without nihilism

The claims of Chen et al. regarding air pollution and life expectancy are somewhat supported by data but far from proven. As noted above, we believe that if the results in question are worth publishing, then they would be worth publishing even if they were not statistically significant at the 5% level. The current rules of publication seem to us to be simultaneously too loose (in the sense of accepting the highly questionable analysis indicated in Figure 1) and too restrictive (in essentially demanding statistical significance, obtained some way or another, as a condition for acceptance).

One might reply that the scientific literature is self-correcting and so we should not worry so much about imperfect or erroneous methods; shaky findings are unlikely to show up on replication. Unfortunately, things do not always work out so well; once researchers know what to expect, they can continue finding it, given all the degrees of freedom available in data processing and analysis (Simmons, Nelson, and Simonsohn, 2011, Gelman and Loken, 2014). As we have written earlier (Gelman, 2013) in the context of a different set of controversial claims, the systematic publication of statistically significant overestimates can lead to "a boom-and-bust cycle of hype and disappointment or, worse, an explaining-away of failed replications if too much trust is placed in the original finding."

And, in the meantime, speculations are presented as fact. For example, the China air pollution study was featured in a *New York Times* article (Wong, 2013) that referred unquestioningly to "the 5.5-year drop in life expectancy in the north," as well as in a *New Yorker* article by a Pulitzer prizewinning reporter (Johnson, 2013) who simply wrote that a study "noted that pollution from coal reduces average life expectancy in northern China by five and a half years," with no indication that the "five and a half years" number was just a point estimate, even setting aside questions about the validity of that estimate.

We need a way of handling such claims that falls between acceptance and dismissal. And we also are glad that Chen et al. produced Figure 1, which made the problems with their study so clear. We would not want criticisms such as ours to serve as a disincentive for authors to display the fit of their models to data. Better for problems to be out in the open than swept under the rug. What the authors did right was to be transparent about their model choices and the implications and then create a plot that made this easy for the reader to digest. But there is an inherent problem with incentives in publication and publicity of research such that the desire to achieve statistically significant results can lead to the acceptance of modeling choices that are supported by neither theory nor data.

Respiration is important. Increasing lifespan is important. There's lots of evidence that air pollution is bad for you. Policies matter. Environmental policies and environmental outcomes can and should be studied by quantitative researchers. Data are never perfect but we still need to move forward. Subject-matter researchers spend decades of their life establishing subject-matter expertise, and they don't always have full statistical expertise. That is fine. There is a division of labor. We are statistics researchers and don't have much subject-matter knowledge about environmental science, even though we publish papers in the area. Regression discontinuity is a great idea. Causal inference from observational data is difficult but still needs to be done. There's often no easy way to control for background variables in a quasi-experiment. These researchers did the best they could. Their conclusions are consistent with much of the literature and their paper was accepted in a leading scientific journal. Even if their work is imperfect it should not be dismissed. We focus on a statistical concern with this paper because statistics is what we do. We suspect that an improved analysis of these data would yield a higher level of uncertainty, leading to 95% intervals that contain zero. This would not mean that the true effect is zero, it just means there is some level of uncertainty in the effect given an analysis based only on the data at hand. We are skeptical about an effect of 5 years of life but we could be wrong. We think it would be fine for the journal to publish an article just like this one, but without the cubic polynomial and with a smaller and non-statistically-significant estimate of the effect.

We have the impression that research journals have an implicit rule that under normal circumstances they will publish this sort of quantitative empirical paper only if it has statistically significant results. That's a regression discontinuity right there, and researchers in various fields (for example, Button et al., 2013) have found evidence that it introduces some endogeneity in the selection variable.

4. References

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

Berger, J., and Pope, D. (2011). Can losing lead to winning? *Management Science* **57**, 817–827.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. Technical report, Department of Economics, University of Michigan.

Chen Y., Ebenstein, A., Greenstone, M., and Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences* **110**, 12936–12941.

Gelman, A. (2009). A statistician's perspective on "Mostly Harmless Econometrics: An Empiricist's Companion," by Joshua D. Angrist and Jorn-Steffen Pischke. *Stata Journal* **9**, 315–320.

Gelman, A. (2013). Is it possible to be an ethicist without being mean to people? *Chance*.

Gelman, A., and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University.

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240.

Johnson, I. (2011). In the air: Discontent grows in China's most polluted cities. *New Yorker*, 2 Dec., 32–37.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* **142**, 67–697.

Lee, D. S., and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* **48**, 281-355.

Pope, C. A., and Dockery, D. W. (2013). Air pollution and life expectancy in China and beyond. *Proceedings of the National Academy of Sciences* **110**, 12861–12862.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359–1366.

Sumner, S. (2013). China life expectancy bleg. The Money Illusion blog, 5 Aug. `http://www.themoneyillusion.com/?p=22764`

Wong, E. (2013). Pollution leads to drop in life span in northern China, research finds. *New York Times*, 9 Jul., A6.