Research Article

Research and Politics
January-March 2015: 1–7

DOI: 10.1177/2053168015569830
rap.sagepub.com
SAGE

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

Andrew Gelman[1] and Adam Zelizer[2]

Abstract

It is common 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 and that arbitrarily attribute variation to the high-degree polynomial or the discontinuity. This paper examines two recent studies that make use of regression discontinuity to discuss evident practical problems with these estimates and how they interact with pathologies of the current system of scientific publication. First, we discuss a recent study that estimates the effect on air pollution and life expectancy of a coal-heating policy in China. The reported effects, based on a third-degree polynomial, are statistically significant but substantively dubious, and are sensitive to model choice. This study is indicative of a category of policy analyses where strong claims are based on weak data and methodologies which permit the researcher wide latitude in presenting estimated treatment effects. We then replicate a procedure from Green et al., in which regression discontinuity is used to recover estimated treatment effects relative to an experimental benchmark, to illustrate one practical problem with the regression discontinuity estimates in the coal-heating paper: high-degree polynomials yield noisy estimates of treatment effects that do not accurately convey uncertainty. We recommend that (a) researchers consider the problems which may result from 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

Regression discontinuity analysis

Regression discontinuity (RD) methods are one of the standard techniques used in statistics and econometrics to obtain causal inference from observational data. But implementations of RD can have serious problems in practice, especially with the common approach of controlling for high-degree polynomials of the underlying continuous predictor. In a companion paper (Gelman and Imbens, 2014) we present evidence that controlling for high-order polynomials in RD analysis results in noisy estimates with poor statistical properties and confidence intervals that are too narrow. In the present paper we discuss evident practical problems with these estimates and how they interact with pathologies of the current system of scientific publication.

We demonstrate with a recent well-publicized example in public health where a high-degree polynomial control in an RD analysis led to implausible conclusions. The magnitude and significance of reported treatment effects were highly sensitive to model specification. We then extend a paper by Green et al. (2009) to illustrate that high-degree polynomial estimates such as those reported in the public health paper are subject to uncertainty and noise not captured by reported p-values. In addition to implying that

[1]Department of Statistics and Department of Political Science, Columbia University, NY, USA
[2]Department of Political Science, Columbia University, NY, USA

Corresponding author:
Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027, USA.
Email: gelman@stat.columbia.edu

researchers should show much more caution with such models, this experience suggests a rethinking of conventional ideas of robustness to model specification.

RD analysis, introduced by Thistlewaite and Campbell (1960), has recently 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.

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, 2011) 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 these 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 for checking the robustness of RD estimates of the treatment effect:

1. "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."
2. "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, for example, by Calonico et al., 2014). The challenge is what to do *after* following this advice.

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

We discuss in the context of a paper by Chen et al. (2013) that received a great deal 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 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 these results are interesting even if the 95% CIs happen to include zero. There is an unfortunate convention that "p less than 0.05" results are publishable while "non-significant" results are not. The life expectancy of 500 million people is important, and it is inappropriate to wait on statistical significance to make policy decisions in this area.

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 of Lee and

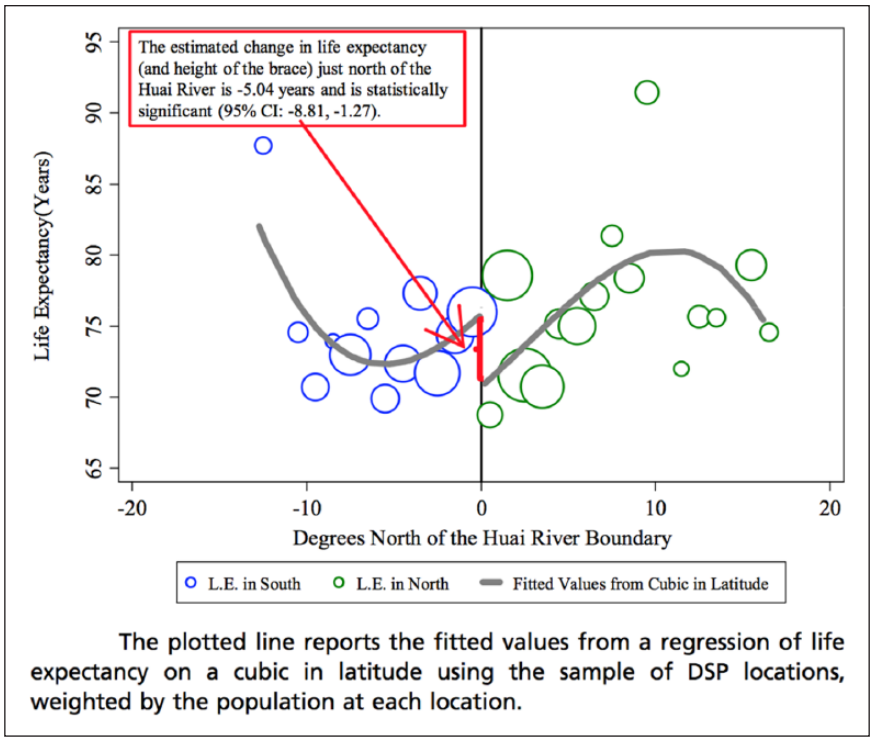

Figure 1. Key graph from Chen et al. (2013) showing their regression discontinuity analysis. Each circle represents the average from a set of locations in China.
CI: confidence interval; LE: life expectancy

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 μg/m^3)	2.89***	2.63***	1.84***	1.95***	1.52**
	(0.56)	(0.49)	(0.63)	(0.59)	(0.72)
Life Expectancy (years)	-1.62	-1.29	-5.52**	-5.67**	-5.43*
	(1.66)	(1.68)	(2.39)	(2.36)	(2.94)

Figure 2. Excerpts of a table from the Supplementary Material online from Chen et al. (2013). Our problem with all these models is that they do not include other predictors and that the residual errors are large (see Figure 1). Thus 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.

Lemieux reported above. However, we are far less than 97.5% sure that the effects are in the direction that the authors claim[1].

Table S.9 in the Supplementary Material online, reproduced in part here as Figure 2, 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. The large, statistically significant, estimated treatment effect at the discontinuity depends on the functional form employed. The higher-degree polynomials have the advantage of being more general but the disadvantage of yielding noisy and often implausible estimates. The implausibility is illustrated in Figure 1; the noise we shall discuss in a bit.

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 model. Our point is rather that the

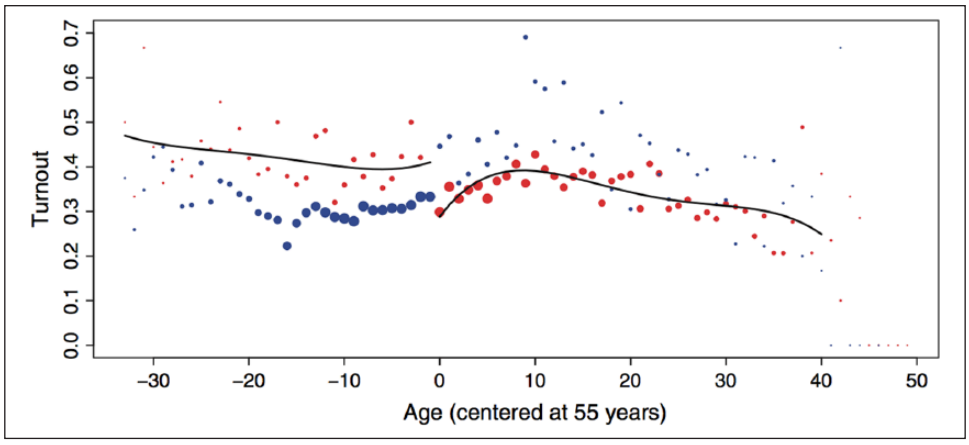

Figure 3. Graph from Green et al. (2009) showing a missing-data construction used to study the performance of regression discontinuity estimates: "The red circles depict average voting rates among observed voters, grouped by year of age, which has been rescaled so that zero (age 55) is the point of discontinuity. The blue circles depict average voting rates among counterfactual voters [who were artificially removed from the dataset]. The red circles to the left of the age cutoff (where age equals 0) represent the treatment group, which received the experimental mailings. The red circles to the right of the cutoff represent the control group, which received no experimental mailings. The size of the circles is proportional to the number of observations in each age group." Reproduced by permission of Oxford University Press.

headline claim, and its statistical significance, is highly dependent on a model choice that may have a data-analytic purpose, but which has no particular scientific basis. Figure 1 indicates to us that neither the linear nor the cubic nor any other polynomial model is appropriate here. Instead, there are other variables not included in the model which distinguish the circles in the graph.[2]

We suggest caution in the estimation and reporting of high-degree polynomial estimates. In the following section, we demonstrate a general, undesirable feature of these estimates, one which may have contributed to the implausible estimates reported above: high-degree polynomials produce noisy estimated treatment effects with standard errors that do not accurately reflect the true degree of uncertainty in the estimate.

Reanalysis of RD models fitted to data from a voter mobilization experiment

The China pollution study illustrates an example where a fitted high-degree polynomial has the effect of adding noise to the estimated discontinuity treatment effect. The standard error reported from a discontinuity regression does not account for systematic error in the fitted model – in this case, the high-degree polynomial – and thus represents a lower bound on uncertainty (Green et al., 2009). Similarly, Gelman and Imbens (2014) demonstrate that RD inferences can have much worse than nominal coverage (for example, p-values of 0.05 occurring more than 10% of the time) in the presence of systematic error in the fitted curve.

It is not always so apparent how noisy the RD estimate can be when a single model is being fitted to data. We demonstrate the problem here in an example of a randomized experiment in which an RD structure is artificially created by removing data. Our analysis elaborates on an example from Green et al. (2009), who take data from a randomized experiment that Gerber et al. (2008) conducted on potential voters. Green et al. create an artificial discontinuity, removing all the treated people in the sample who were below the age of 55 and removing all the controls who were 55 and older. In their 2009 study, Green et al. had the full data from the original randomized experiment, which they used to estimate a benchmark experimental treatment effect, and a partial dataset with a discontinuity, for which they used RD methods to estimate the effect at age 55. Figure 3 shows the observed (in red) and missing (in blue) data for this analysis.

The broken line in Figure 3 shows a fitted fourth-degree polynomial with a discontinuity at age 55; the fitted curve looks reasonable (if perhaps a bit too sharply sloped just on the right-hand side of the breakpoint) and the estimated jump at the discontinuity yields a treatment effect estimate that is consistent with that obtained from the experimental benchmark.

But the reasonability of this high-degree polynomial estimate depends on the breakpoint chosen, as we can see by repeating the discontinuity analysis at a range of potential cutpoints from ages 40 to 70.

Figure 4 shows the RD estimates as a function of the age discontinuity, along with 95% error bounds from the estimated regressions. These graphs do *not* display fitted curves; rather, each point on each graph shows the coefficient estimate and uncertainty from a single RD model.

The RD estimates are very noisy and in several places highly misleading (for example, at ages 43, 46, and 66). The estimate at age 55 happens to look good but this seems to be largely a matter of luck. For many age cutoffs, the corresponding fitted models are noisy and implausible with

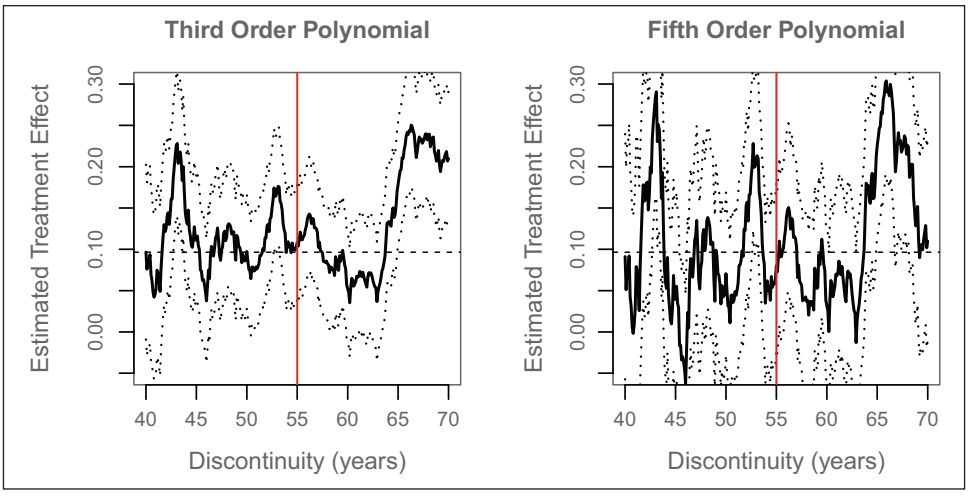

Figure 4. Regression discontinuity estimates of the effects of the voter mobilization treatment, based on replicating the procedure of Green et al. (2009) separately, for each age cutoff from 40 to 70. At each age, the graphs show the estimate and 95% confidence interval from the fitted regression, controlling for a third or fifth-order polynomial. The red bar shows the result at the age cutoff of 55, which was used by Green et al. as illustrated in Figure 3, and the horizontal dotted line at 0.097 corresponds to the estimated average treatment effect from the full data from the randomized experiment. The regression discontinuity estimates are noisy, especially when controlling for the fifth-degree polynomial.

dramatic up and down swings, chasing the data in a manner reminiscent of Figure 1.[3]

The point here is not that RD with high-order polynomials *always* gives bad answers. Rather, without some constraint on the smoothness of the fitted functions, we do not recommend simply fitting such a model. Including high-order polynomials is not a universally conservative approach. These models can fail, generating noisy estimated discontinuity coefficients with p-values that do not reflect the uncertainty of the model specification.

Discussion

Publication of speculative findings on particulate pollution and life expectancy

Our goal in reassessing the findings of Chen et al. is to call into question the way scholars control for higher-order polynomials in RD analysis and the substantive implications that follow from these data-analytic decisions. We are not saying that particulate matter does not kill, that this topic should not be studied, or that these findings should not be published in a high-profile journal. The accompanying article by Pope and Dockery (2013) considers why the conclusions reached by Chen et al. might be scientifically plausible.

Rather, we see that example as indicative of a category of policy analyses where strong claims are based on weak data, with high-order polynomial RD designs one example of how researchers can amplify the magnitude or significance of estimated treatment effects with an eye toward publication. Researchers are not alone in placing too much value on statistically significant findings or those with large substantive effects. What we suggest is a two-step: that authors retreat from strongly model-based claims of statistical significance and that journals accept that non-statistically-significant findings on important topics are still worth publishing.

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

At a technical level, we understand the appeal of controlling for high-order polynomials of the assignment variable, following the general principle that it is safest and most conservative to control for potential confounders to reduce bias.

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, or Figure 4, which shows the large differences in the magnitude of estimated treatment effects across models.

In well-designed RD studies, the underlying predictive effect of the assignment variable is clear. For example, in the Lee (2008) 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. 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 criticizing the use of high-degree polynomials in RD adjustments, we are not recommending global linear adjustments as an alternative. In some settings a linear relationship can make sense (for example in data with a simple before–after structure), but in general our concerns about systematic error will not disappear with the use of a simpler form. What we *are* warning against is the appealing but misguided view that users can correct for arbitrary dependence on the forcing variable by simply including several polynomial terms in a regression. We recommend that any RD analysis include a plot such as Figure 1 showing data and the fitted model, and that users be wary of any resulting inferences based on fits that don't make substantive sense. Our message is also consistent with that of Green et al. (2009), who recommend comparing observational studies with controlled experiments where possible.

Skepticism without nihilism

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 (Gelman and Loken, 2014; Simmons et al., 2011). As we have written earlier (Gelman, 2013b) 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 – those that are provocative and substantively important while falling short of conventional levels of statistical significance – that falls between acceptance and dismissal. 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. The authors were quite correctly transparent about their model choices and the implications of these choices, and they created a plot that made the data and model easy for the reader to digest. Regardless of this good-faith effort, there remains an inherent problem with incentives in publication and publicity of research: the desire to achieve statistically significant results can lead to the acceptance of modeling choices that are supported by neither theory nor data.

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 is a discontinuity right there, and researchers in various fields (for example, Button et al., 2013) have found evidence that it introduces endogeneity in the forcing variable.

Acknowledgements

We thank Shigeo Hirano, Jennifer Hill, Eric Voeten, and two reviewers for helpful comments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Some of the material on the China air pollution example appeared earlier in blog form (Gelman, 2013a).

Funding

This work was supported by the National Science Foundation and Institute for Education Sciences (grants CNS-1205516 and DE R305D140059).

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Supplementary material

The online appendix is available at: http://isps.yale.edu/research/data/d016#.VI-9jSvF-Sp Citation

Notes

1. Recall that 97.5% is the posterior probability of a positive effect given $p = 0.05$, under a flat prior. 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% CI represent very large effects that seem a priori unlikely.
2. The strong upward slope of the model at the discontinuity is curious. We would expect a negative slope if life expectancy decreased as a function of ambient pollution. The fitted model also implies that moving from five to 12 degrees south of the Huai River boundary is associated with a 10 year

increase in life expectancy. A more plausible explanation is that the outlier is explained by omitted variables.

3. The RD analysis is only intended to recover a local treatment effect and so, to be fair, we should compare not with an average treatment effect but rather with a local average treatment effect, as estimated from the full data from the randomized experiment. Such an analysis yields an average treatment effect that is stable at around 0.0965 (that is, a nearly 10 percentage point increase in voter turnout) for most of the range of ages, with some evidence that the effect rises to around 0.15 above age 65.

References

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

Button KS, Ioannidis JPA, Mokrysz C, et al. (2013) Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5): 365–376.

Calonico S, Cattaneo MD and Titiunik R (2014) Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica* 82(6): 2295–2326.

Chen Y, Ebenstein A, Greenstone M, et al. (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(32): 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(2): 315–320.

Gelman A (2013a) Evidence on the impact of sustained use of polynomial regression on causal inference (a claim that coal heating is reducing lifespan by 5 years for half a billion people). In: Statistical modeling, causal inference, and social science blog, 5 August. Available at: http://andrewgelman.com/2013/08/05/evidence-on-the-impact-of-sustained-use-of-polynomial-regression-on-causal-inference-a-claim-that-coal-heating-is-reducing-lifespan-by-5-years-forhalf-a-billion-people/ (accessed 10 February 2015).

Gelman A (2013b) Ethics and statistics: Is it possible to be an ethicist without being mean to people? *Chance* 26(4): 52–55.

Gelman A and Imbens G (2014) Why high-order polynomials should not be used in regression discontinuity designs (No. w20405). Cambridge, MA: National Bureau of Economic Research.

Gelman A and Loken E (2014). The statistical crisis in science. *American Scientist* 102(6): 460.

Gerber AS, Green DP and Larimer CW (2008) Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review* 102(01): 33–48.

Green DP, Leong TY, Kern HL, et al. (2009) Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis* 17(4): 400–417. Replication materials at ISPS Data Archive, available at: http://isps.yale.edu/research/data/d016.VI-9jSvF-Sp (accessed 6 November 2014).

Hill J (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1): 217–240.

Johnson I (2013) In the air: Discontent grows in China's most polluted cities. *New Yorker*, 2 December 2013, pp.32–37.

Lee DS (2008) Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 142(2): 675–697.

Lee DS and Lemieux T (2010) Regression discontinuity designs in economics. *Journal of Economic Literature* 48: 281–355.

Pope CA and Dockery DW (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(11): 1359–1366.

Thistlewaite D and Campbell D (1960) Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology* 51(6): 309–317.

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