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.