# Criticism as asynchronous collaboration: An example from social science research*

Andrew Gelman

23 Feb 2022

**Abstract**

I discuss a published paper in political science that made a claim that aroused skepticism. The reanalysis is an example of how we, as consumers as well as producers of science, can engage with published work. This can be viewed as a sort of collaboration performed implicitly between the authors of a published paper and later researchers who want to understand or use the published work.

## 1. Introduction

Collaboration is essential to statistics: applied problems motivate the development and evaluation of new methods, which in turn allow new applied problems to be solved. This cycling between concerns of statistics and subject matter will only work in the presence of strong connections between statistics and applied fields.

The present article illustrates a different sort of collaboration, performed implicitly between the authors of a published paper and later researchers who want to understand or use the published work. This is not a collaboration in the usual sense, as it does not require that the different research groups ever contact each other—but it is arguably the most important, and presumably the most common, form of research connection. Just about anyone who publishes a paper with a statistical result has a hope that various complete strangers will read it and take its findings seriously enough to attempt to understand and replicate them.

The most usual forms of scholarly engagement with published work are citation, appreciation, and refutation. But these do not fully capture the experience of reading and evaluating a paper, in particular the uncertainties involved in balancing theoretical and evidential claims.

Here I discuss a published paper in political science (Barfort, Klemmensen, and Larsen, 2021) that made a claim—"politicians winning a close election live 5–10 years longer than candidates who lose"—that aroused my skepticism. ***My purpose here is not to formally rebut that paper (although I do present several arguments against it) but rather to use this as an example of how we, as consumers as well as producers of science, can engage with published work.*** Fortunately in this case the data from the published study were available to all for reanalysis. The authors of the paper under discussion were provided an additional opportunity to respond and had nothing formal to add beyond what they had already written, but they did supply some corrections which were incorporated into the present article.

## 2. Reactions to a published paper

The paper under discussion states: "we exploit a regression discontinuity design with unique data on the longevity of candidates for US gubernatorial office. The results show that politicians winning a close election live 5–10 years longer than candidates who lose." Their analytic approach and results are well summarized by Figure 1.

*To appear in *Stat.* I thank Erik Gahner Larsen, Bill Harris, and two anonymous reviewers for helpful comments and the U.S. Office of Naval Research, Institute for Education Sciences, and Sloan Foundation for partial support of this work.

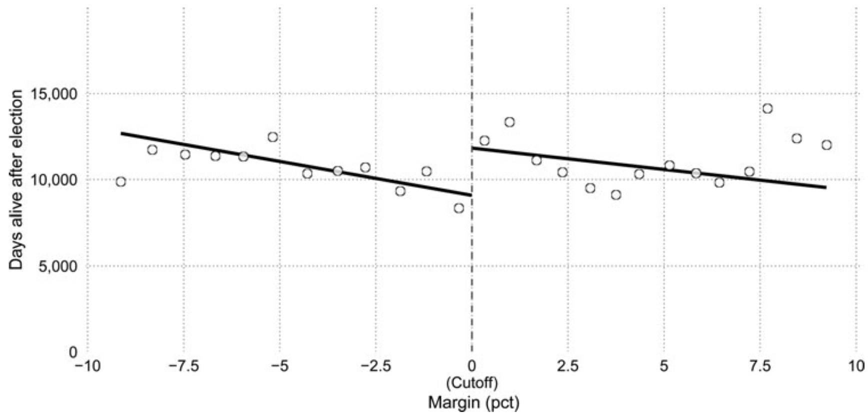


Figure 1: *Published graph showing the fitted regression discontinuity model of governors' elections and lifespans, titled, "The causal effect of winning on longevity."*

This regression discontinuity analysis makes use of the insight that, if treatment assignment (in this case, a candidate losing or winning the election) is entirely dependent on the running variable (in this case, the difference in the candidates' vote shares), then there is no possibility of selection bias: in econometric or statistical jargon, the assignment is exogenous or ignorable conditional on this variable. The challenge is that, by design, there is imbalance and zero overlap of treatment and control groups with respect to the running variable, hence the problem is typically framed in terms of statistical modeling and robustness, as the inference for the causal effect will depend crucially on the method used to adjust for this imbalance.

Despite the $p < 0.05$ statistical significance of the discontinuity in the above graph, and despite the assurance of the authors that this finding is robust, no, I do not believe that winning a close election causes U.S. governors to live 5–10 years longer.

How can I say this? I will answer in five ways:

1. Common sense
2. Our experience as consumers of research
3. Statistical analysis
4. Statistical design
5. Sociology of science.

Our discussion of these five reasons will shed some light, I hope, on the ways in which authors and researchers interact.

**Common sense.** Five to ten years is a huge amount of lifetime. Even if you imagine dramatic causal sequences (for example, you lose an election and become depressed and slide into alcoholism and then drive your car into a tree), it's hard to imagine such a huge effect. For statistical reasons discussed below, it is not surprising that this sort of study can come up with an effect size estimate that is implausibly large, even in the absence of any true effect of that magnitude. For now, though, I will just say that such a large effect, while not physically impossible, gives us reason for doubt.

**Our experience as consumers of research.** The recent history of social and behavioral science is littered with published papers that made claims of implausibly large effects, supported by statistically significant comparisons and seemingly solid research methods, which did not hold up under replication or methodological scrutiny. Some examples that I happened to have looked into in some detail include the claim that beautiful parents were eight percentage points more likely to have girls, the claim that students at Cornell had extra-sensory perception, the claim that women were three times more likely to wear red or pink clothing during certain times of the month, the claim that single women were twenty percentage points more likely to support Barack Obama during certain times of the month, and the claim that political moderates perceived the shades of gray more accurately than extremists on the left and right. We've been burned before. We've been burned enough times that we realize we don't have to follow the now-retired dictum of Kahneman (2011) that "you have no choice but to accept that the major conclusions of these studies are true." For a particularly vivid example, I recommend the story of Nosek, Spies, and Motyl (2013) of a theoretically founded, statistically significant result they found in one of their research studies—which then did not show up in an attempted replication. It can be as easy to fool oneself as to fool others, and we know by now that confidence in the part of authors and publication in a respectable journal is not enough. It is not that nothing should be trusted; rather, the point here is that published results in the human sciences are often mistaken. In particular, common methods of statistical analyses routinely lead to overstatement of evidence.

**Statistical analysis.** If there truly is no such large effect that losing an election causing you to lose 5 to 10 years of life, then how could these researchers have found a comparison that was (a) statistically significant, and (b) robust to model perturbations? My quick answer is researcher degrees of freedom or forking paths (Simmons, Nelson, and Simonsohn, 2011, Gelman and Loken, 2014): the many different ways the analyses could have been done, contingent on data. Figure 1 might look compelling, but now consider Figure 2, which shows the raw data (for all elections between 1945 and 2012 that were decided by less than 10 percentage points, just counting politicians who are no longer living). For ease of interpretation I've plotted the data in years rather than days.

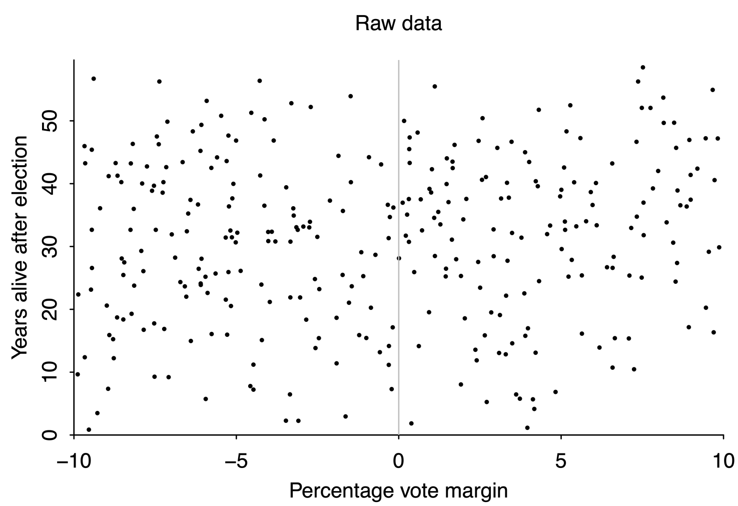


Figure 2: *Raw data from the governors' elections and lifespans analysis. Compare to the aggregates plotted in Figure 1.*

Next we throw a local linear smoother (loess) on there and see what we get; this is shown in Figure 3. And Figure 4 shows three loess fits, one for negative $x$ and one for positive $x$, at which point a discontinuity opens up.

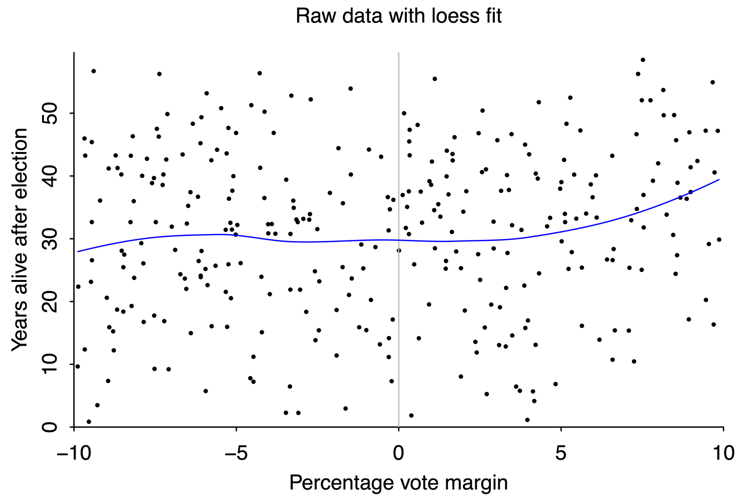


Figure 3: *Data from the governors' elections and lifespans analysis along with a fitted local linear smoother.*

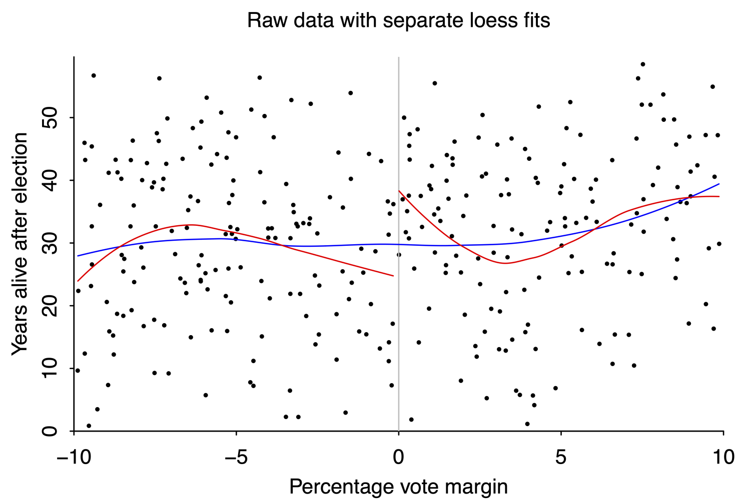


Figure 4: *Data from the governors' elections and lifespans analysis along with the fitted local linear smoother (blue curve) and two separate locally linear smoothers fit separately to the data where the candidates lost and won the election (red curves). A discontinuity at $x = 0$ counteracts a strong and implausible negative relation between vote margin and lifespan near the boundary.*

We've seen this movie before (for example, Gelman and Zelizer, 2015). When the data are relatively flat near the break point, but the fitted continuous curve from the regression discontinuity model happens to have a steep negative slope right near the discontinuity, then a positive jump is needed to counter an essentially random pattern in the overfitted curves.

The fit is noisy, and finding the discontinuity all depends on there being this strong negative relation between future lifespan and vote margin in this one election—but just for vote margins in this $\pm 5$ percentage point range. Without that negative slope, the discontinuity goes away.

At this point you might say, no, the authors actually fit a local linear regression, so we can't blame the curve, and that their results were robust. We'll get to that. My first point here is that the data are super-noisy, and fitting different models to these data will give you much different

results. Again, remember that it makes sense that the data are noisy—there's *no good reason at all* to expect a strong relationship between vote margin in an election and the number of years that someone will live afterward. Indeed, from any usual way of looking at things, it's ludicrous to think that a candidate's life expectancy is:
30 years if he loses an election by 5 percentage points
25 years if he loses narrowly
35 years if he wins narrowly
30 years if he wins by 5 percentage points.
It's a lot more believable that this variation is just noise, some artifact of the few hundred cases in this dataset, than that it represents some general truth about elections, or even about elections for governor.

As a starting point, we perform a linear regression, following the data choices of the paper under discussion by restricting to elections between 1945 and 2012 that were decided by less than 10% of the vote and excluding all missing data, including politicians who were still alive at the time of the construction of the dataset. Figure 5 shows the result: an estimated effect is 2.4 years with a standard error of 2.4 years, i.e., consistent with noise.

```
                   coef.est coef.se
(Intercept)        78.60     4.05
won                 2.39     2.44
age                -0.98     0.08
decades_since_1950 -0.21     0.51
margin             -0.11     0.22
---
n = 311, k = 5
residual sd = 10.73, R-Squared = 0.35
```

Figure 5: *Simple regression estimating effect on lifespan of winning an election for governor, adjusting for age at time of the election, epoch, and victory margin.*

But what about the robustness as reported in the published article? My answer to that is, first, the result is not so robust, as is indicated by the above graph and regression and demonstrated further in the next section—and I wasn't trying to make the result go away, I was just trying to replicate what they were doing in that paper,—and, second, as Simonsohn (2016) explains, "Robustness checks involve reporting alternative specifications that test the same hypothesis. Because the problem is with the hypothesis, the problem is not addressed with robustness checks." Simonsohn illustrates that point with an amusing example in which he roots through the General Social Survey to find a (spurious) relationship between the response to a horoscope-reading question and the serial number that was randomly assigned to survey respondents. His statistically significant result survives a reasonable-looking robustness check.

To return to the general theme of remote collaboration: the authors of a published paper supply robustness checks as part of an asynchronous dialogue with readers who might be skeptical of the claims. Alternative analyses are part of any good statistical study, but generally the point should be to explore and understand the limitations of one's conclusions, not to rule out alternative explanations.

**Statistical design.** Another way to understand my skepticism is to consider the design of this sort of study. A priori we might consider an effect size of one additional year of life to be large, and on the border of plausibility. But this study has cannot reliably detect effects this small. You can see that from the standard errors on the regression. If an estimate of 5–10 years is two or three standard errors from zero, than an effect of 1 year, or even 2 years, is statistically undetectable. So the study is really set up only to catch artifacts or noise. This is what Button et al. (2012) call "power failure."

**Sociology of science.** If you are a social scientist, statistical methods should be your servant, not your master. It's tempting to say that the authors of the paper in question followed the rules of regression discontinuity analysis and that it's not fair to pick on them. Indeed, in other work, we have criticized the use of high-degree polynomials in discontinuity analyses (Gelman and Imbens, 2019), so you might even say that by using linear regression, the authors were heeding my own advice. But the point of social science is not to follow rules, it's to gain understanding and make decisions. When following the rules gives silly results, it's time to look more carefully at the rules and think hard about what went wrong. That's how many in the field of psychology responded (appropriately, in my opinion) to the replication crisis of the 2010s.

Yes, it's possible that I'm wrong. Perhaps losing a close election really did kill these candidates for governor. It's possible. But I don't think so. I think this paper is just another example of statistical rule-following that's out of control.

## 3. Working with the data

I followed the link at the published article and downloaded the data. The code didn't quite run as is—the R code required a `.csv` file but all I could find was a `.tab` file, so I changed the code accordingly. Then when I ran the next bit of the code:

```
if (sha1(df_rdd) != "300cc29bbecd2b630016c9bd2c8ef958dcc1b45d"){
  error("Wrong data file loaded or data has been changed!") }
```

I got an error:

```
Error in error("Wrong data file loaded or data has been changed!") :
  could not find function "error"
```

So I checked by typing:

```
sha1(df_rdd)
```

and got this:

```
"e26cc6acbed7a7144cd2886eb997d4ae262cf400"
```

So maybe the data did get changed. I have no idea. But I was able to reproduce the graphs so things were probably OK. I include these details not because they are important in themselves but as a small example of the operational details that rarely appear in print.

I also noticed some data problems, such as cases where the politician's death date came decades before his election. There were a bunch of these, for example John Quincy Adams was listed as having run for governor of Massachusetts five times between 1867 and 1871 but as having died in 1826. I'm not trying to say that this is a horrible dataset. This is just what happens when you try to replicate someone's analyses. Problems come up.

Once I started to go through the data, I realized there are a lot of analysis choices. The article in question focuses on details of the discontinuity analysis, but that's really the least of it, considering that we have no real reason to think that the vote margin should be predictive of longevity. The most obvious thing is that the best predictor of how long you'll live in the future is how long you've lived so far. So I included current age as a linear predictor in the model. It then would be natural to interact the win/loss indicator with age. Another issue is what to do with candidates who are still living. The published analysis just discarded them. It might make sense to include such data using survival analysis. I didn't do this, though, as it would take additional work. I guess that was the motivation of the authors, too. No shame in that—I cut corners with missing data all the time—but it is yet another researcher degree of freedom.

Another big concern is candidates who run in multiple elections. It's not appropriate to use a regression model assuming independent errors when you have the same outcome for many cases. To keep things simple, I just kept the first election in the dataset for each candidate, but that's not the only choice; instead you might, for example, use the average vote margin for all the elections where the candidate ran.

Finally, for the analysis itself: it seems that the published regressions were performed using an automatic regression discontinuity package, `rdrobust`. I'm sure this is fine for what it is, but, again, I think that in this observational study the vote margin is not the most important thing to adjust for. Sure, you want to adjust for it, as there's no overlap on this variable, but we simply don't have enough cases right near the margin (the zone where we might legitimately approximate win or loss as a randomly applied treatment) to rely on the discontinuity alone. 400 or so cases is just not enough to estimate a realistic effect size here. To keep some control over the analyses, I just fit some simple regressions. It would be possible to include local bandwidths etc. if that were desired, but, again, that's not really the point. The discontinuity assignment is relevant to our analysis, but it's not the only thing going on, and we have to keep our eye on the big picture if we want to seriously estimate the treatment effect.

I then did some manipulations to check the dates in the file and clean the data, counting each candidate just once. My code was ugly. I did check the results a bit, but I wouldn't be surprised if I introduced a few bugs. I created a `decades_since_1950` variable as I had the idea that longevity might be increasing, and I put it in decades rather than years to get a more interpretable coefficient. I restricted the data to 1945–2012 and to candidates who were no longer alive at this time because that's what was done in the paper, and I considered election margins of less than 10 percentage points because that's what they showed in their graph, and also this did seem like a reasonable boundary for close elections that could've gone either way (so that we could consider it as a randomly assigned treatment).

Then I fit some regressions, starting with the one shown in Figure 5, with each time the outcome being the number of additional years lived by the candidate after the election. We start by following the standard advice to interact the treatment with the most important predictor, which in this case is age; the result appears in Figure 6.

```
                  coef.est coef.se
(Intercept)       78.81     5.41
won                1.92     8.25
age               -0.98     0.10
decades_since_1950 -0.21    0.51
margin            -0.11     0.22
won:age            0.01     0.16
---
n = 311, k = 6
residual sd = 10.75, R-Squared = 0.35
```

Figure 6: *Regression estimating effect on lifespan of winning an election for governor, also including the interaction between victory and age at time of the election, following up on the model in Figure 5 by including the interaction of treatment with the most important predictor.*

The coefficient for the interaction with age is so small and so noisy that it does not seem helpful, so we return to the earlier regression in Figure 5 which yielded an estimated treatment effect of 2.4 years (not 5–10 years) with standard error 2.4.

What about excluding `decades_since_1950`? It can make sense to remove a predictor whose estimate is so noisy. Figure 7 shows the result: essentially no change from that shown in Figure 5.

```
            coef.est coef.se
(Intercept) 78.62     4.05
won          2.36     2.44
age         -0.99     0.08
margin      -0.11     0.22
---
n = 311, k = 4
residual sd = 10.72, R-Squared = 0.35
```

Figure 7: *Regression estimating effect on lifespan of winning an election for governor, removing the predictor for epoch. Compare to Figure 5.*

We could exclude age, which I wouldn't recommend given how strong a predictor it is, but we can try it; the result appears in Figure 8. Unfortunately, now the estimate's even smaller and noisier. We should include age in the model in any case.

```
            coef.est coef.se
(Intercept) 30.14     1.75
won          1.70     3.01
margin       0.10     0.27
---
n = 311, k = 3
residual sd = 13.25, R-Squared = 0.01
```

Figure 8: *Regression estimating effect on lifespan of winning an election for governor, removing the predictor for age. Compare to Figure 5.*

We could increase the precision by including all elections where the vote margin was within 20 points; see Figure 9. Now we have almost 500 cases, but we're still not seeing that large and statistically significant effect.

```
                   coef.est coef.se
(Intercept)        76.24     3.29
won                 1.00     1.83
age                -0.93     0.06
decades_since_1950 -0.18     0.39
margin              0.02     0.09
---
n = 497, k = 5
residual sd = 10.83, R-Squared = 0.33
```

Figure 9: *Regression estimating effect on lifespan of winning an election for governor, including all elections where the winning margin was within 20% of the vote. Compare to Figure 5.*

I'm not claiming that my regressions are absolutely better than the ones in the published paper. I'm pretty sure they're better in some ways and worse in others. I just found it surprisingly difficult to reproduce their results using conventional approaches. I believe they found what they found, but I call the result fragile, not robust.

Still not sure what to try, I re-ran the analysis going back to keeping only the elections with margins under 10 points, but including the duplicates as separate observations in the regression. Figure 10 shows the result. The coefficient for winning the election is almost statistically significant with a $z$-score of 1.7, but still not in that 5–10 year range. Also, as we've already discussed, it doesn't make sense to count a politician as multiple data points, as each person only has one life.

```
                   coef.est coef.se
(Intercept)        74.65     3.18
won                 3.15     1.89
age                -0.91     0.06
decades_since_1950 -0.03     0.41
margin             -0.19     0.17
---
n = 499, k = 5
residual sd = 10.93, R-Squared = 0.33
```

Figure 10: *Regression estimating effect on lifespan of winning an election for governor, including duplicate cases where a politician ran for governor more than once and thus was included multiple times in the dataset. Compare to Figure 5.*

I thought I could get cute and remove `age` and `decades_since_1950` and maybe something like the published paper's result would appear, but no luck, as can be seen in Figure 11.

```
            coef.est coef.se
(Intercept) 28.45     1.32
won          2.84     2.30
margin      -0.12     0.20
---
n = 499, k = 3
residual sd = 13.32, R-Squared = 0.00
```

Figure 11: *Regression estimating effect on lifespan of winning an election for governor, removing predictors for age and epoch. Again, the hoped-for statistically significant effect fails to appear.*

I then reproduced the authors' results using the `rdrobust` package in R, which yielded an estimated effect of 7.5 years with standard error 2.4. I tried to match this by re-running the basic linear model but just in the zone where the candidates were within 5 percentage points of winning or losing, which yielded the promising result shown in Figure 12.

```
                   coef.est coef.se
(Intercept)        71.03     6.87
won                 7.54     3.76
age                -0.90     0.12
decades_since_1950 -0.36     0.78
margin             -1.14     0.68
---
n = 153, k = 5
residual sd = 11.46, R-Squared = 0.31
```

Figure 12: *Regression estimating effect on lifespan of winning an election for governor, restricting to elections where the winning margin was within 5% of the vote. Compare to Figure 5.*

We're getting closer: the estimate is 5–10 years, and it's 2 standard errors from zero. We can juice it up a bit more by removing `decades_since_1950` (a reasonable choice to remove) and age.

We should really keep age in the model, no question about it—not including age in a remaining-length-of-life model is about as bad as not including smoking in a cancer model—but we remove it just to see what happens. The result appears in Figure 13 and, we now have an estimated effect of more than 10 years of life that is 2.8 standard errors from zero.

```
lm(formula = more_years ~ won + margin, data = data, subset = subset4)
            coef.est coef.se
(Intercept) 22.70     2.52
won         11.92     4.31
margin      -2.00     0.77
---
n = 153, k = 3
residual sd = 13.34, R-Squared = 0.05
```

Figure 13: *Regression estimating effect on lifespan of winning an election for governor, restricting to elections where the winning margin was within 5% of the vote and excluding predictors for age and epoch. Compare to Figure 12.*

Alternatively can start with the regression discontinuity model in `rdrobust` as used in the published paper and keeping only the first race for each candidate, which yields an estimated effect of 4.1 years with standard error 2.8. If we then include age as a predictor in the `rdrobust` call, we get an estimate of 2.4 with standard error 1.9. The robust setting in this package turns out not to matter much in this example—but the analysis is sensitive to the bandwidth (yes, the bandwidth is estimated from the data, but that just tells us it can be noisy; the fact that something is calculated automatically using some theory and a computer program doesn't mean it's correct in any particular example) and to the decision of how to handle candidates with multiple elections in the dataset, and to the modeling of the age predictor.

## 4. Discussion

### 4.1. No smoking gun

It's easy to criticize research when the forking paths are all out in the open (as with the Bem, 2011, study of ESP) or when statistics show that your sample size is too low to detect anything by a factor of 100 (as in the Kanazawa, 2007, study of beauty and sex ratio; see Gelman and Weakliem, 2009) or when there are obvious forking paths and failed replications (as in many notorious studies in social psychology) or when almost all the data have been excluded from the analysis (as in the study by Kim, Zhang, and Zhong, 2021, of unionization and stock prices; see Gelman, 2019) or when there's flat-out research misconduct.

This example discussed here is a bit different. It's a clean analysis with clean data. The data are even publicly available (which allowed me to make the above graphs), and the researchers responded directly to my concerns. But, remember, honesty and transparency are not enough. If you do a study of an effect that is small and highly variable (which this one is: to the extent that winning or losing can have large effects on your lifespan, the effect will surely vary a lot from person to person), you've set yourself up for scientific failure: you're working with noise.

I'm not happy about this, but that's just how quantitative social science works. So let me emphasize again that a study can be fatally flawed just by being noisy, even if there's no other obvious flaw in the study.

Or, to put it another way, there's an attitude that causal identification + statistical significance = discovery, or that causal identification + robust statistical significance = discovery. But that attitude is mistaken. Even if you're an honest and well-meaning researcher who has followed principles of open science.

### 4.2. Awkwardness of this exercise

This was all a lot of work. I did it because it's worth doing some work now and then—but don't forget the larger point, which is that I was suspicious from the start because of the implausibly large effect sizes and knew ahead of time not to assume that causal identification + statistical significance + robustness tests = discovery, because we've been burned on that combination many times before.

To put it another way, *don't* give the claims of a published or preprinted social science study the benefit of the doubt, just because someone like me didn't bother to go to all the trouble to explain its problems. The point here is not to pick on the authors of this particular study—not at all. Their work is admirably open and clear.

Again, there is nothing special about regression discontinuity here. All the same concerns would arise with any observational study, whether it uses matching, difference in differences, nonparametric modeling, instrumental variables, synthetic controls, or just plain regression. The same problems arise in experiments too, what with issues of missing data and extrapolation. Indeed, sometimes it's worse with designed experiments, because of the illusion that they give clean causal identification. With all these kinds of study, if the underlying effect size is small and your measurements are noisy, you're drawing dead (as they say in poker). This can happen in any study, ranging from a clean randomized experiment at one extreme, to a pure observational study on the other.

### 4.3. Potential rebuttal

As noted, authors and different groups of readers collaborate implicitly and asynchronously, and so it can help to move forward by anticipating objections to one's work. To this end, I constructed the following response by a hypothetical skeptic of my skepticism:

> *Hey, destructive statistician. Stop with your nihilism. The authors of this peer-reviewed paper did a standard, reasonable analysis and found a large, statistically significant, and robust effect. All you did was perform a bunch of bad analyses. Even your bad analyses uniformly found positive effects. They just weren't statistically significant, but that's just because you threw away data and used crude methods. It's well known that if you throw away data and use inefficient statistical analysis, that your standard errors tend to increase. In summary, you have maligned a competent, high-quality paper—one that passed peer review—by the simple expedient of re-running their analysis with fewer data points and a noisier statistical method and then concluding that their results were not statistically significant.*

The above response sounds kind of reasonable—indeed, readers might agree with it—and it has the form of a logical argument, but I think it's wrong, for several reasons.

First, regarding the point that the coefficient shows up positive in all my analyses, hence making this nothing more than a dispute over statistical significance: This where it's helpful to have a Bayesian perspective—or, if you'd prefer, a design-based perspective. If the true effect is in the range of 1 year of life (in either direction), and this is studied with data and analysis that yield an estimate with standard errors of 2 years or more, then the signal will be dominated by noise, and a positive coefficient can be explained as just the summation of some highly variable numbers that happened to end up in one direction or another.

Second, regarding the idea that I've replaced their state-of-the-art regression discontinuity analysis with various noisy linear regressions: Actually, no. The robust adaptive discontinuity regression does not have more statistical efficiency than the linear model. These are just different models. Indeed, the decision of the authors of the original paper to not include age as a predictor *lowers* the efficiency of the analysis. Regarding sample size: I took out the duplicate elections featuring the same candidate because it's inappropriate to consider these as independent data points; indeed, including them in that way can just give you artificially low standard errors. (It should be easy enough to show this using a bootstrap analysis or something like that.) The real point, though, is you can't tell much about the efficiency of a study by looking at the $z$-score or statistical significance of a coefficient estimate.

### 4.4. Reactions of the original authors

It is good to anticipate criticism but better to actually receive it. In this case, the authors of the paper under discussion had contacted me directly, and after I posted my reactions online, one of them responded (Larsen, 2020), and I posted a response to their response (Gelman, 2020). We had some back-and-forth about the details of how I attempted to reproduce the published analysis, but the most important sticking point was the plausibility of the effect size. In an email correspondence, the author wrote:

> "If the 'true' effect is, say, .5 years, we would need a lot more data to detect that. In other words, I believe your point 4 is spot on. That's indisputable and simply a limitation of the paper. I am not sure I would say that the paper is 'fatally flawed' for that reason, but I get the point. Are you saying that we might as well have detected a negative effect of 5–10 years? (I mean, if we're simply playing around with noise here.) . . . We had a lot of discussions about the effect size when writing up the paper and, to be honest, I don't know what's a realistic/common sence effect size to expect here. When you work on a paper long enough, your results become 'common sense' (or to paraphrase Duncan Watts . . . 'everything is obvious once you know the answer')."

I replied:

> "I suspect any true effect would be much less than 5 years, so that a pattern in the data that shows up in the regression is basically just noise. I think that in a replication study of other data with the same sample size, for example from some state legislatures or whatever, it would be possible to get an estimate of −5 yrs or −10 yrs. I expect that a fully preregistered study would not yield a statistically significant result, but your study was not preregistered (nor are most of mine. I'm not holding my own research up as a model here) . . . Is an effect size of 5 or 10 years reasonable? I agree that we can't know, but it really doesn't make sense to me, especially given that candidates typically face many elections in their lifetimes. The point of the type M errors is that, with such a small sample and such noisy data, anything statistically significant estimate you will find will *have* to be huge. So, in that sense, I don't see your estimate of 5–10 yrs as providing any useful information regarding the magnitude of the effect: Even it turns out that there is some average effect, even as large as 1 or 2 years (which I would really doubt), the observed estimate of 5–10 yrs is an artifact of the small sample and high variability."

The author added:

> "Also, a good friend of mine informed me (when he saw the study) that we're not the first to show an interest in this subject. Apparently, a couple of economists published a study on the topic in 2019 as well. They find a positive effect but a lot smaller, i.e. 3 years when looking at elections post-1908. Interestingly, they find that governors running in elections post-1908 (the sample most similar to ours) live $\sim$ 6 years longer (Panel B, Column 1–2, Table 3). I am not linking to this study to make a point about the results (again, I believe you have a good point on the fact that we would need more statistical precision to estimate a smaller effect), but it might be of interest to cite that study as well to make a broader point about inference and robustness?"

I took a look at that other study (Borgschulte and Vogler, 2019), which considers legislative as well as gubernatorial elections. Figure 14 shows their key plot. Their is strikingly similar to the main result of the paper under discussion (see the Figure 1); again, a noisy fit yielded a strong negative slope on the two sides of be boundary, which then allows a positive jump at the discontinuity. The point is that regression discontinuity analyses have sufficient researcher degrees of freedom to allow statistically significant coefficients to be extracted from data that are consistent with pure noise, as discussed by Gelman (2017). And this returns us to my skepticism about a mechanism by which losing elections would reduce candidates lifespans by 5 or more years on average.

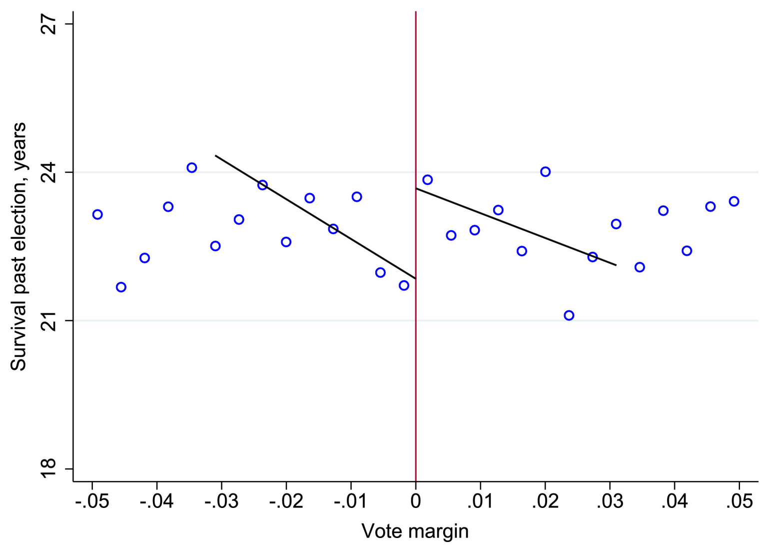


Figure 14: *Summary plot from Borgschulte and Vogler (2019) showing a different regression discontinuity analysis of lifespans for candidates for elective offices. As with Figure 1, which is based on similar data, a positive jump at the discontinuity appears as the result of allowing a strong negative slope at the two sides of the boundary.*

### 4.5. Reading and using research articles as a form of asynchronous collaboration

When a research paper is successful, we can assume that most of its readers probably have no direct connection with the authors. And the afterlife of a published paper is not just with its readers but also with later researchers who apply its findings or methods. As such, post-publication criticism is an important part of the feedback mechanism of science, and we can see it as a form of collaboration occurring across time as well as space.

Direct, or synchronous, collaboration is characterized by frequent "handshaking": theoretically redundant calculations or observations that allow the different members of a research team to check their answers against each other. This kind of confidence-building is more challenging in the more general setting when there is no contact between the original researchers and the replicators. The present article gives a sense of how this process can go.

## References


Barfort, S., Klemmensen, R., and Larsen, E. G. (2021). Longevity returns to political office. *Political Science Research and Methods* **9**, 658–664.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* **100**, 407–425.

Borgschulte, M., and Vogler, J. (2019). Run for your life? The effect of close elections on the life expectancy of politicians. *Journal of Economic Behavior & Organization* **167**, 18–32.

Gelman, A. (2017). Forking paths plus lack of theory = No reason to believe any of this. `https://statmodeling.stat.columbia.edu/2017/12/29/forking-paths-plus-lack-theory-no-reason-believe/`

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* **44**, 16–23.

Gelman, A. (2019). Another regression discontinuity disaster and what can we learn from it. `https://statmodeling.stat.columbia.edu/2019/06/25/another-regression-discontinuity-disaster-and-what-can-we-learn-from-it/`

Gelman, A. (2020). Comment. `https://statmodeling.stat.columbia.edu/2020/07/02/no-i-dont-believe-that-claim-based-on-regression-discontinuity-analysis-that/#comment-1373309`

Gelman, A., and Imbens, G. (2017). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics* **37**, 225–456.

Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 460–465.

Gelman, A., and Weakliem, D. (2009). Of beauty, sex, and power: Statistical challenges in estimating small effects. *American Scientist* **97**, 310–316.

Gelman, A., and Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research and Politics* **2**, 1–7.

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology* **244**, 133–140.

Kim, J. B., Zhang, E. X., and Zhong, K. (2021). Does unionization affect the manager-shareholder conflict? Evidence from firm-specific stock price crash risk. *Journal of Corporate Finance* **69**, 101991.

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* **7**, 615–631.

Larsen, E. G. (2020). A response to Andrew Gelman. `https://erikgahner.dk/2020/a-response-to-andrew-gelman/`

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.