Article

Research and Politics
January-March 2018: 1–9

DOI: 10.1177/2053168018756668
journals.sagepub.com/home/rap
SAGE

Visual heuristics for marginal effects plots

Thomas B. Pepinsky

Abstract

Common visual heuristics used to interpret marginal effects plots are susceptible to Type-1 error. This susceptibility varies as a function of (a) sample size, (b) stochastic error in the true data generating process, and (c) the relative size of the main effects of the causal variable versus the moderator. I discuss simple alternatives to these standard visual heuristics that may improve inference and do not depend on regression parameters.

Keywords

Interaction terms, marginal effects plots, conditional hypotheses, data visualization

Introduction

The interpretation of interaction terms in political science is a topic of wide interest (Brambor et al., 2006; Braumoeller, 2004; Berry et al., 2012; Esarey and Sumner, forthcoming; Hainmueller et al., 2017; Kam and Franzese, 2007). An influential article by Brambor et al. (2006), in particular, has transformed how political scientists study and interpret interactive hypotheses.[1] In addition to reminding researchers that they must include both constitutive terms and interaction terms if they wish to test interactive hypotheses, the authors write that "The analyst cannot … infer whether X has a meaningful conditional effect on Y from the magnitude and significance of the coefficient on the interaction term either… It means that one cannot determine whether a model should include an interaction term simply by looking at the significance of the coefficient on the interaction term" (p. 74). The authors propose instead to use "marginal effects plots" to calculate the estimated marginal effect of the variable of interest across substantively meaningful values of the moderating variable.

Marginal effects plots have since become ubiquitous in political science. Despite their ubiquity, there is little analysis of their performance as a tool for identifying interactive effects. Two recent papers have begun to look more closely at the marginal effects plot. Hainmueller et al. (2017) show that marginal effects plots (and indeed, any hypothesis test relying on an interaction term) rely on the assumption that the effect of the causal variable of interest is linear and constant across the values of the moderating variable. By contrast, Esarey and Sumner (forthcoming) argue that marginal effects plots usually have inappropriate coverage because of the problem of multiple comparisons. My contribution in this manuscript is to draw attention to the visual heuristics that researchers implicitly use when they interpret marginal effects plots.

In this article I demonstrate that applied researchers have drawn incorrect conclusions from Brambor et al. (2006). The appropriate test for the presence of linear interaction effects *is* given by the significance of the coefficient on the interaction term. Commonly used visual heuristics, which I identify below, will often fail compared to this test. Marginal effects plots have other uses, but they should not be used to test for the presence of linear interaction effects.

Focusing on Type-1 error, or the problem of false positives, I ask the following question: *How frequently will visual inspection of a marginal effects plot suggest that interaction effects exist when the true data generating process is not interactive?* To investigate, I generate simulated data for a binary treatment variable D and an additional predictor X. The true data generating process is $Y = \alpha + \beta_1 D + \beta_2 X + \epsilon$, where $\alpha = \beta_1 = \beta_2 = 1$. Here, X does not moderate the effect of D on Y. I then used a

Department of Government, Cornell University, USA

Corresponding author:
Thomas B. Pepinsky, Department of Government, Cornell University, 322 White Hall, Ithaca, NY 14853, USA.
Email: pepinsky@cornell.edu

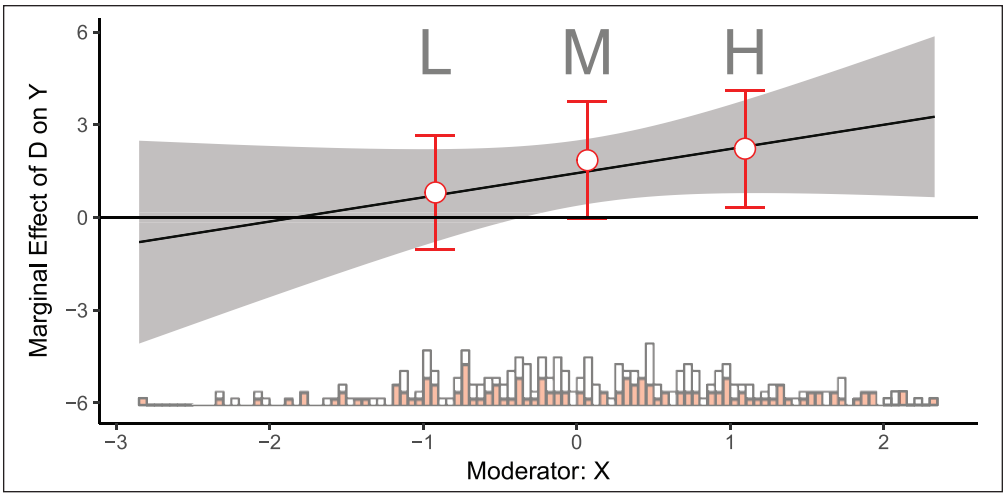

Figure 1. The marginal effect of D on Y.

marginal effects plot to test whether the effect of D varies across values of X by estimating the regression $Y = a + b_1 D + b_2 X + b_3 X \cdot D$. I use the inter.binning command in the interflex package in Hainmueller et al. (2017), which produces marginal effects plots as well as what they term a "binning estimator" that allows for the effect of D on Y to be nonlinear in X. The results are in Figure 1.

Visual inspection of Figure 1 suggests that the effect of D on Y is positive at high values of X, and is indistinguishable from zero at low values of X. (The same conclusion emerges from a visual inspection of the binning estimator as well.) In practice, it is common for empirical researchers to conclude from marginal effects plots of this sort that the effect of D on Y depends on the value of X, and to reject the null hypothesis that there is no interactive effect between X and D.[2]

That conclusion is incorrect. The coefficient on the interaction term, b_3, *is the* test of whether the effect of D on Y depends on the value of X (see Kam and Franzese, 2007: 50). In the above example, we know the true data generating process, so we know that the effect of D on Y does not depend on the value of X. And indeed, the p-value on b_3 is 0.147, which does not to reject the null hypothesis that the effect of D does not depend on X at the 95% confidence level. In the remainder of this article I clarify what a marginal effects plot tests and how this is different from the hypothesis that the effect of D on Y varies by X; illustrate the consequences of using marginal effects plots to test the latter hypothesis; recapitulate standard recommendations about how to test interactive hypotheses; and propose a more informative marginal effects plot that discourages their misuse.

Learning from marginal effects plots

Marginal effects plots contain two pieces of information. The first is the slope of the "marginal effect line," which is determined by the coefficient b_3. The second is the width of the confidence intervals, which depend on the estimated variances and covariances between b_1 and b_3. From these pieces of visual information, the researcher makes inferences about the data generating process using heuristics: is the slope positive or negative, is the slope "large" or "small," does the line cross zero, and for what range of values do its confidence intervals include zero?

When there is no interactive effect, the true value of b_3 is zero. However, estimating a regression with an interaction term will produce non-zero estimates of b_3; *in expectation* these estimates will be zero, but in any application they will almost always indicate a non-zero slope for the marginal effects line. Regression with an interaction term will likewise produce non-zero estimates of the variance-covariance matrix, including the covariance of the non-existing interaction effect and the main effects.

Figure 2 illustrates how such a marginal effects plot *ought* to look when there is no interaction between X and D.

Because the effect of D does not depend on X, the line is flat across values of X, and the confidence intervals are bounded away from zero. To generate such a clean visual result, however, I had to set the sample size to 100 000 and the variance of ϵ to 1.

When visual results are not so clean, researchers commonly follow one of two visual heuristics. The first, which I term the "crosses zero" heuristic, looks to see whether or not the confidence intervals capture zero for some portion of the range of X and do not for some other range of X. If so, the inference is then that *the effect of D on Y is nonzero for some range of X, and zero elsewhere.* The second, which I term the "compare extremes" heuristic, looks to see whether there is overlap across the entire range of the confidence band. If not, the inference is then that *the effect of D on Y differs across the values of X*. The range of the confidence band depends on the range of values of X

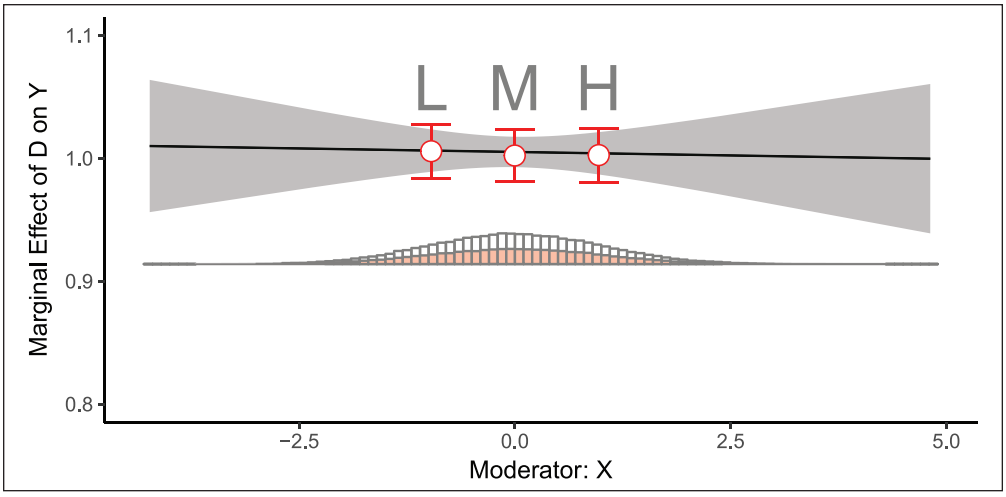

Figure 2. The marginal effect of D on Y.

across which marginal effects are calculated; for the purposes of this discussion, I consider the relevant range to be the observed range of X in the sample.[3] To be clear, these two heuristics amount to tests of different hypotheses. The compare extremes estimator is already subject to critique because if the extremes of the marginal effects plot include values that lie beyond the area of common support, then inferences are particularly fragile (see e.g., Hainmueller et al., 2017: 3). To my knowledge, the crosses zero heuristic has never been identified, but is widely used.

One other piece of information that may test whether or not an interaction effect exists is the coefficient b_3. If b_3 is small and statistically indistinguishable from zero, this would be evidence against the hypothesis that the marginal effect of D depends on X. However, researchers often hold that the coefficient b_3 is not a test of whether interaction effects exist, because that coefficient does not provide information about the marginal effect D on Y at various levels of X, which is usually the quantity of interest. Both the crosses zero heuristic and the compare extremes heuristic may be interpreted as attempts to better study the marginal effects of D across values of X.

We are left with what seems to be an impasse. The coefficient on the interaction term is not a meaningful test of the marginal effect D on Y at various levels of X. And yet it proved very easy to develop a case where standard heuristics based on a marginal effects plot produced misleading conclusions that interaction effects do exist. The solution is to recognize that the following two statements are entirely consistent with one another.

1. The coefficient of b_3 expresses the term $\frac{\partial(\partial Y / \partial D)}{\partial X}$, or "does the effect of D differ across the values of X?"
2. The confidence intervals around the point estimates across values of X, as found in a marginal effects plot, are the value-by-value correct confidence intervals for each of those conditional effects of D on Y. Each expresses the term $\partial Y / \partial D \mid X$, or "what is the effect of D on Y when $X = x$?"

Differences in $\partial Y / \partial D \mid X$ across values of X cannot be easily translated into evidence about whether the effects of D depend on X. For present purposes, however, the key is that the crosses zero heuristic *does not* generally translate the latter into the former. An extensive review of the specific hypotheses tested by various interaction models can be found in Kam and Franzese (2007), pp. 43–92.

How frequently do researchers employ the crosses zero heuristic? I consulted each of the articles replicated by Hainmueller et al. (2017) and checked for evidence that authors explicitly based their inferences on a marginal effects plot *rather than* on the statistical significance of the interaction term. The authors argue that this sample represents "high profile" articles that likely took "special care to employ and interpret these models correctly." By my count, 7 out of 22—or nearly one out of every three articles—fulfill this criterion.[4] In the majority of the remaining 15 cases, the coefficient on the interaction term was itself significant, obviating the need to choose one or the other. For the same reasons that Hainmueller et al. (2017) argue that their replications represent a lower bound on the true rate of problematic multiplicative interaction terms, my count may also represent a lower bound on how often visual heuristics are used to identify interaction effects.

Simulations

The preceding discussion explains why it is not correct to compare marginal effects to make inferences about the *presence of* interactive effects. To illustrate the dangers of doing so, I use simulations. Based on the data generating

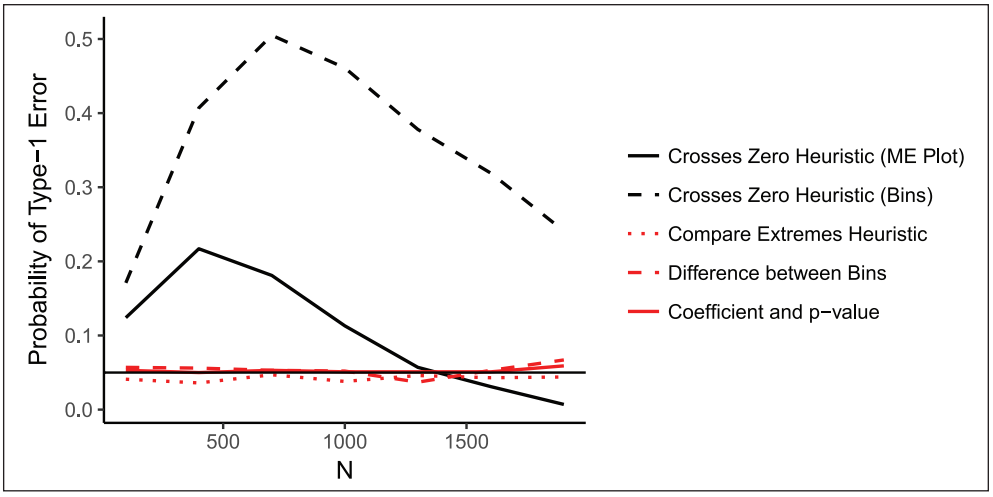

Figure 3. Type-1 error rates for four different heuristics.

process outlined in the introduction, I created 1000 simulated datasets and created "virtual" marginal effects plots for each. I then implemented five tests: three based on the heuristics outlined above, one based on the coefficients from the binning estimator, and one based on the coefficient on b_3.

1. *Crosses zero heuristic* If the estimated marginal effect of D on Y is both statistically distinguishable from zero across at least 25% of the range of X and indistinguishable from zero across at least 25% of the range of X—analogous to Figure 2—I conclude that the marginal effects plot is consistent with the presence of an interactive effect. I implement this test by checking if either of the following conditions hold: the confidence interval of the 25th (75th) percentile of X captures zero, the confidence interval of the 75th (25th) percentile of X excludes zero, and the confidence interval of the 97.5th (2.5th) percentile of X excludes zero.[5] Note that this implementation of the crosses zero heuristic is fairly conservative in requiring statistical insignificance across at least an entire quartile of X. If I had decreased this requirement to a quintile or a decile, my conclusions would be stronger.
2. *Crosses zero heuristic (bins)* This heuristic applies the same logic of the crosses zero heuristic to a plot derived from the binning estimator. If the confidence interval of the low (high) tercile captures zero, and the confidence interval of the high (low) tercile does not capture zero, and the point estimate for each tercile falls in the order *(Low, Middle) > High* or *(High, Middle) > Low*, then I conclude that the binning estimator plot is consistent with the presence of an interactive effect.
3. *Compare extremes heuristic* If the maximum value of the lower confidence interval is greater than the minimum value of the upper confidence interval, then I conclude that that marginal effects plot is consistent with the presence of an interactive effect. Recognizing the critiques that exist of this heuristic, note here that I study only cases where the confidence band extends to the observed maximum and minimum of a normally distributed moderator whose values are independent of the causal variable.
4. *Differences between bins* If the two-sided p-value for a test of the equality of the first and third bins is less than .05, I conclude that the binning estimates are consistent with the presence of an interactive effect.
5. *Coefficient and p-value* If the p-value associated with b_3 is less than .05, then I conclude that a standard regression-based approach is consistent with the presence of an interactive effect.

I then repeat this process hundreds of times, varying four parameters: the sample size n, the variance of ϵ, the ratio of β_1 to β_2, and α. The results appear below.

First, I fix $V(\epsilon) = 4, \beta_1 = 1, \beta_2 = 1$, and $\alpha = 1$, and then vary sample size from $n = 100$ to 2000. There is clear evidence that with a sample size of 1000 or less, the crosses zero heuristic based on a marginal effects plot is overconfident relative to regression coefficients about the presence of an interactive effect. The performance of the same visual heuristic applied to the binning estimator is even worse. The performance of regression coefficients, a formal test of the differences between bins, and the compare extremes heuristic are all invariant to sample size.

The small sample performance of the cross-zero heuristic in Figure 3 is noteworthy because it runs counter to common expectations that small samples lead to

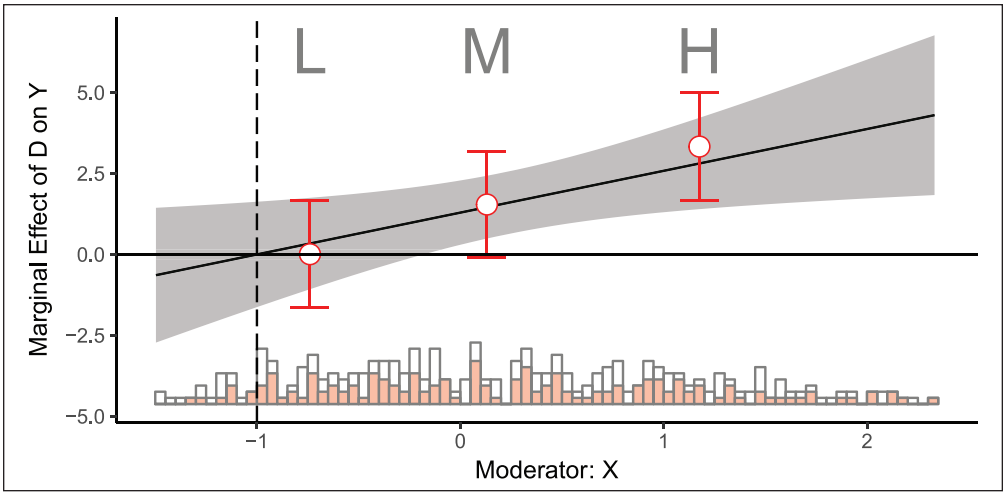

Figure 4. A linear interaction effect.

conservative tests that are more likely to fail to reject the null when an alternative hypothesis is true. In identifying interaction effects, the crosses zero heuristic is anticonservative in small samples.

In the Appendix I vary other features of the simulations. Specifically, I vary the unexplained variance in the model ($V(\epsilon)$), the ratio of β_1 and β_2, and the value of α. Taken together, the results provide further evidence of how the crosses zero heuristic increases the likelihood of Type-1 error.

Discussion

The crosses zero heuristic is overconfident when interpreted to be a test of the hypothesis that the effect of D varies across the range of X because it is sometimes statistically distinguishable from zero. That overconfidence, moreover, depends on features of the regression such as sample size, the relative size of β_1 to β_2, and model error. The binning estimator is particularly useful for detecting nonlinear interaction effects, but if researchers apply the same visual heuristic when interpreting the plots derived from the binning estimator, they will be even more prone to uncover false interactive effects. On the other hand, the power of both the compare extremes heuristic and the coefficient on the interaction term is that their performance does not depend on sample size, stochastic variance, or the size of the causal effect of interest. The problem is that they themselves are not meaningful tests of any substantive hypothesis unless both X and D happen to be binary. Might it be preferable, then, to condition any inferences on the statistical significance of the interaction term? Knowing the answer depends on not just Type-1 error rates, but also Type-2 error rates.

To explore this, I adjust the data generating process to $Y = \alpha + \beta_1 D + \beta_2 X + \beta_3 X \cdot D + \epsilon$, where $\alpha = \beta_1 = \beta_2 = \beta_3 = 1$. In this case, the true marginal effect of D on Y is 0 when $X = -1$; in the simulations below, I truncate the distribution of X at -1.5 to reflect a situation where the effect of D on Y is zero at the lowest values of X, and positive at higher values of X. Marginal effects plots are appropriate in this case because the effects of D on Y are constant in X by construction. An example appears in Figure 4, with $X = -1$ highlighted.

I then test the performance of each of the five heuristics. To "stack the deck" in favor of the crosses zero heuristic, I only require that the confidence interval includes zero at $X = -1$, and that it excludes zero at the 75th percentile of X. In these simulations, I fix $V(\epsilon) = 4, \beta_1 = \beta_2 = \beta_3 = 1$, and $\alpha = 1$.

Figure 5 shows that with small sample sizes, all five heuristics are likely to fail to reject the null hypothesis that there is no interaction effect when one does exist. As sample size increases, all five heuristics improve, but the crosses zero heuristic based on the marginal effects plot improves the fastest. The crosses zero heuristic applied to the binning estimator is acceptable but too conservative, even with large samples. The formal test of the differences between bins only approaches the performance of the other four heuristics when the sample size is large. These results suggest that marginal effects plots are better suited than coefficients and p-values for identifying interaction effects, but only when we know that these effects exist and the sample size is relatively small. Similar conclusions may be drawn from simulations that increase the ratio of stochastic variance to systematic variance.

Finally, I consider a case where the effects of D on Y are nonlinear in X. Specifically, I investigate a data generating process with the following form:

- where $X < -.5, Y = \alpha + \beta X + \epsilon$; and
- where $X \geq -.5$, $Y = \alpha + 2 * (X + .5)^2 * D + \beta X + \epsilon$.

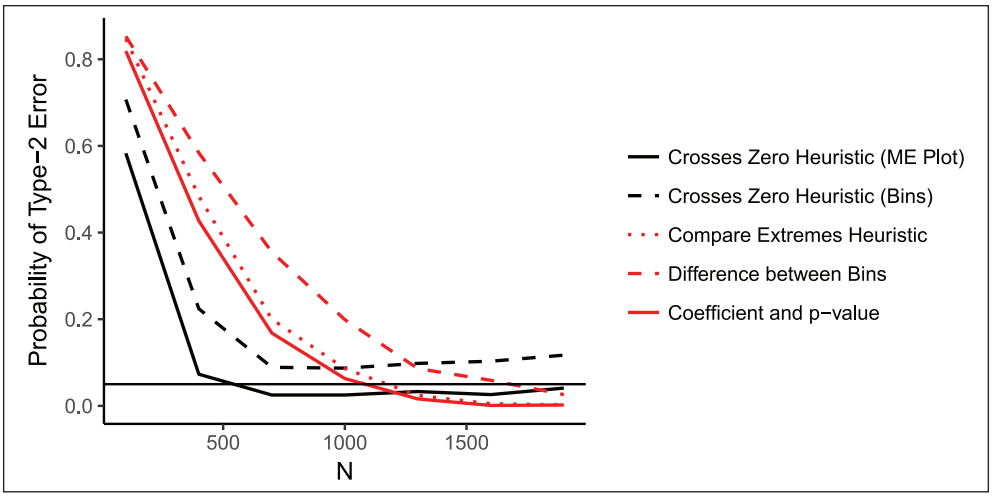

Figure 5. Type-2 error rates for four different heuristics.

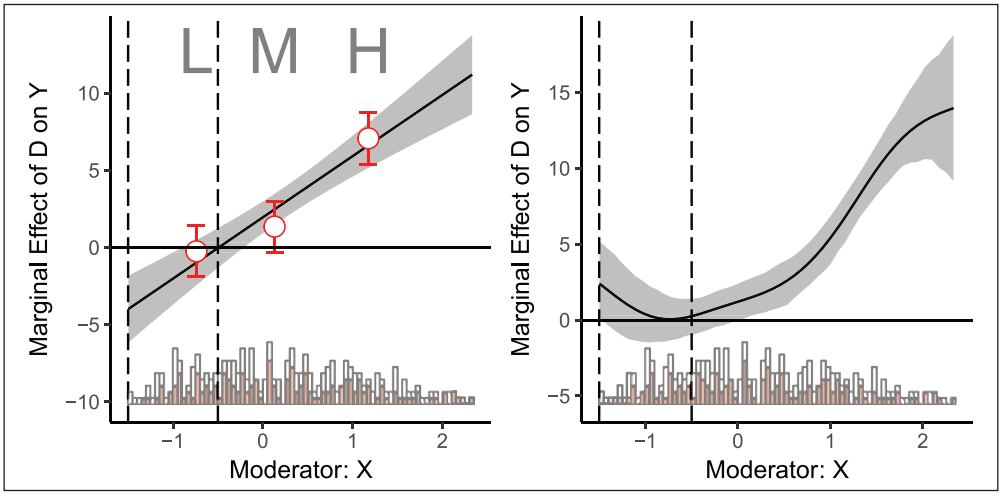

Figure 6. A nonlinear interaction effect.

Here D has no effect on Y when $X < -0.5$, but when $X \geq -0.5$ the effect of D increases as a nonlinear function of X itself—the effect of D is small when X is small, and large when X is large. In Figure 6, I present both the standard marginal effects plot and the kernel estimator plots from Hainmueller et al. (2017), highlighting the range of the data $(-1.5 < X < -0.5)$ where the effect of D is zero.

Not surprisingly, the kernel estimator captures the nonlinear effect of D on Y better than does the marginal effects plot, which indicates a negative and statistically significant marginal effect for D at the low range of X. I then test the performance of two of the five heuristics in capturing the "true" interactive effect in Figure 7.[6] In these simulations, I fix $V(\epsilon) = 4$ and $\alpha = 1$.

In these simulations, the crosses zero heuristic nearly always fails to identify the correct nonlinear effect of D on Y in the marginal effects plot, because it shows that the marginal effect of D on Y is *negative and significant* at low values of X. The binning estimator, on the other hand, has almost a 95% chance of detecting the true nonlinear relationship between D and Y.

Recommendations

This article has shown that visual heuristics used to interpret marginal effects plots can lead to misleading substantive conclusions. When there is no interaction between D and X, the crosses zero heuristic is likely to identify a relationship that does not exist. Relative to a simple inspection of the coefficient on the interaction term, marginal effects plots are thus overconfident. When linear interaction effects *do* exist, marginal effect plots accurately capture the substantive quantities of interest.[7]

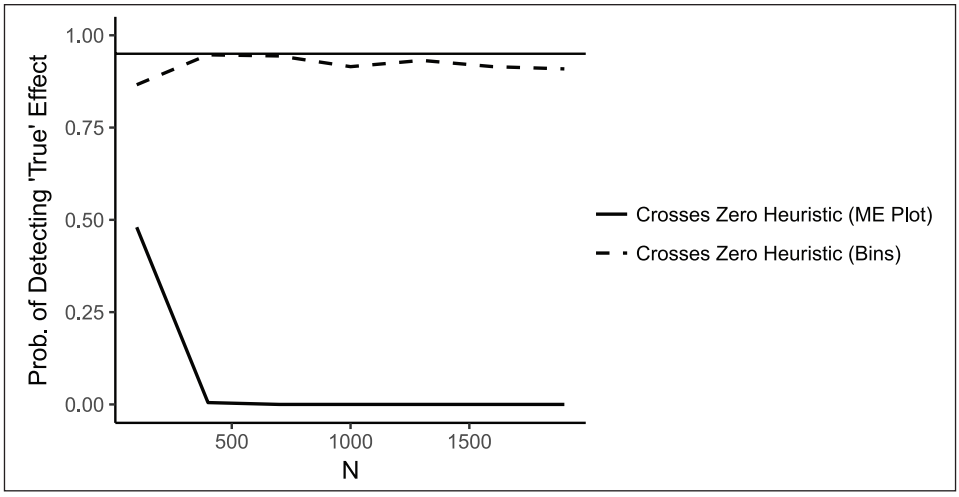

Figure 7. Detecting nonlinear interactions for two different heuristics.

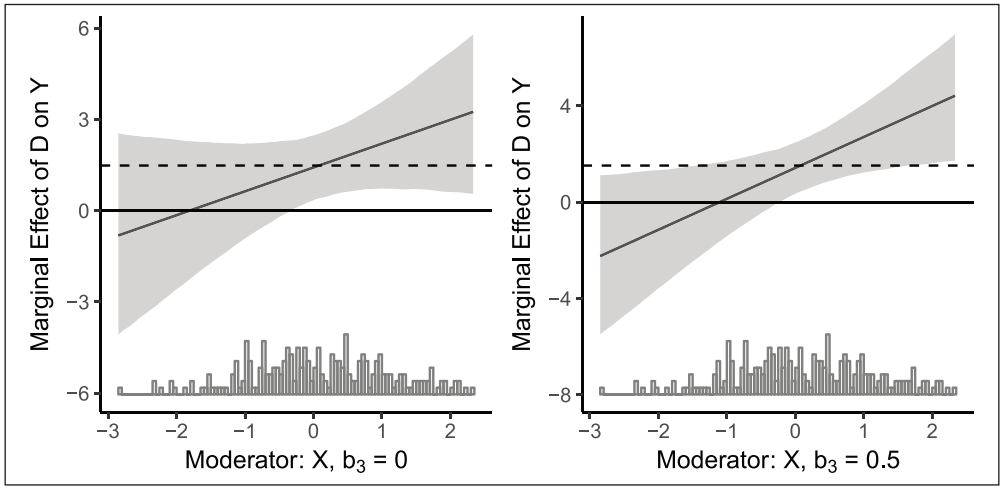

Figure 8. The marginal effect of D on Y.

Brambor et al.'s (2006) most important contribution—amplified by Braumoeller (2004) and Kam and Franzese (2007) in ways that have fundamentally changed research practice—is to shift researchers away from simple inspection of coefficients and standard errors when examining substantive interaction effects. However, Brambor et al.'s (2006) argument that "one cannot determine whether a model should include an interaction term simply by looking at the significance of the coefficient on the interaction term" is incorrect if interpreted to mean that the coefficient on the interaction term does not test whether the effect of D differs across the values of X. Using interaction plots to test for the presence of interactive effects is a mistake.

This discussion suggests some simple guidelines for applied researchers. Assuming linear interaction effects, a conservative strategy that minimizes Type-1 error and which does not depend on sample size, stochastic error, or the relative size of the causal effect of interest would be to only use coefficients and p-values to test for the presence of interaction effects. Although marginal effects plots do calculate the correct marginal effects and their confidence intervals, they do not test for the presence or absence of an interactive effect. Marginal effects plots should be used, then, only to calculate substantive quantities of interest. They are also useful, in combination with histograms of the distribution of the moderating variable, to explore the sensitivity of interaction models to the range of the moderating variable.

Another strategy to improve standard visual heuristics is to add a second reference line that corresponds to the marginal effect of D evaluated at the median of X, as in Figure 8. This line exploits the properties of the compare extremes heuristic, which I demonstrated above to perform about as well as do tests of the significance of the interaction term.

This additional dotted line focuses the eye not only on whether the confidence band includes zero, but also on

whether the entire confidence band spans a common value. The figure on the left plots the same model as in Figure 1, and clearly reveals no interaction effect when $b_3 = 0$. The figure on the right, generated with $b_3 = 0.5$, reveals the appropriate interactive relationship. In combination with the histogram at the bottom of each plot, it is possible as well to inspect whether inferences depend on a few extreme values of X. If so, the methods proposed by Hainmueller et al. (2017) are particularly useful.

I provide open source software in R to create figures similar to Figure 8 in the R package interplot.medline, which is based on the interplot package in R by Solt and Hu (2016).[8] This simple addition to the standard marginal effects plot should discourage researchers from inferring that interaction effects exist when they do not.

Acknowledgements

Thanks to Bryce Corrigan, Justin Esarey, Jens Hainmueller, and anonymous referees for useful comments on previous drafts. I am responsible for all errors.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplementary materials

The supplementary files are available at http://journals.sagepub.com/doi/suppl/10.1177/2053168018756668.

Notes

1. At the time of writing, Google Scholar counts 3781 citations for Brambor et al. (2006), and another 699 for Braumoeller (2004), who presented a similar argument at about the same time.
2. Here is the relevant text from Brambor et al. (2006), who discuss a nearly identical plot: "Confidence intervals around the line allow us to determine the conditions under which presidential elections have a statistically significant effect on the number of electoral parties—they have a statistically significant effect whenever the upper and lower bounds of the confidence interval are both above (or below) the zero line. It is easy to see that temporally proximate presidential elections have a strong reductive effect on the number of electoral parties when there are few presidential candidates. As predicted, this reductive effect declines as the number of presidential candidates increases. Once there are more than 2.9 effective presidential candidates, presidential elections no longer have a significant reductive impact on legislative fragmentation," p. 76.
3. This is the default option in many software packages that automate the creation of marginal effects plots, such as Hainmueller et al.'s (2017) interflex and Solt and Hu's (2016) interplot.
4. Here are two examples. First, from Hellwig and Samuels (2007), pp. 293–294: "It is not possible to infer the extent to which globalization affects electoral accountability directly from the unconditional coefficients and standard errors from the interaction models in Table 1. Although the coefficient on the Economy × Capital Flows term is not statistically significant, the reported standard error pertains only to two specific combinations of values: the marginal effect of Economy when Capital Flows equals 0 or the marginal effect of Capital Flows when Economy equals 0. Figures 1 and 2 better illustrate the degree to which exposure to the global economy conditions the effect of economic performance on election… Once exposure to international trade exceeds 77% of GDP, the positive effects of economy on incumbent vote are no longer statistically significant, as shown by the 95% confidence-interval bands." Next, from Bodea and Hicks (2015), p. 278: "In Model 9, we find a statistically significant and correctly signed (positive) CBI index, with a further positive coefficient of the interaction of CBI and Polity. Interaction terms and their components are, however, difficult to interpret, and Brambor et al. (2006) prescribe that for multiplicative interaction models, inference should be done with meaningful marginal effects and standard errors to determine the conditions under which key variables have a statistically significant effect. We plot the marginal effect of CBI as democracy increases in non-OECD countries in Figure 1(a). The marginal effect is significant only at high levels of Polity, supporting Hypothesis 2.1."
5. This third condition ensures that the interaction effect looks "directional" rather than being essentially flat with the confidence intervals excluding zero only at the midpoint of the range of *X*.
6. The compare extremes heuristic, the difference between bins test, and the coefficients and p-values always indicate an interactive effect, but these are by design not informative of whether the effect of D is zero in some range of X and positive elsewhere.
7. Hainmueller et al.'s (2017) binning estimator fares best in the presence of nonlinear interaction effects, which is the exact purpose for which it was proposed.
8. The package may be installed directly from GitHub following instructions available at https://tompepinsky.com/research/code/

Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

References

Berry W, Golder M and Milton D (2012) Improving tests of theories positing interaction. *Journal of Politics* 74: 653–71.

Bodea C and Hicks R (2015) International finance and central bank independence: Institutional diffusion and the flow and cost of capital. *Journal of Politics* 77 (1): 268–84.

Brambor T, Clark WR and Golder M (2006) Understanding interaction models: Improving empirical analyses. *Political Analysis* 14: 63–82.

Braumoeller BF (2004) Hypothesis testing and multiplicative interaction terms. *International Organization* 58: 807–20.

Esarey J and Sumner JL (forthcoming) Marginal effects in interaction models: Determining and controlling the false positive rate. *Comparative Political Studies*.

Hainmueller J, Mummolo J and Xu Y (2017) How much should we trust estimates from multiplicative interaction models? *Simple tools to improve empirical practice*. Available at: https://ssrn.com/abstract=2739221 or http://dx.doi.org/10.2139/ssrn.2739221

Hellwig T and Samuels D (2007) Voting in open economies: The electoral consequences of globalization. *Comparative Political Studies* 40 (3): 283–306.

Kam CD and Franzese RJ (2007) *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: University of Michigan Press.

Solt F and Hu Y (2016) *Interplot: Plot the Coefficients of Variables in Interaction Terms*. The Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/web/packages/interplot/interplot.pdf (accessed 12 February 2018).