This draft: 2023 11 17
Latest draft: http://urisohn.com/45

Don't Bin, GAM Instead: Hainmueller et al.'s Binning & Kernel Estimators Are Only Valid for Experimental Data

Uri Simonsohn
ESADE Business School
urisohn@gmail.com

Abstract:
Hainmueller, Mummolo, and Xu (2019) identified two problems with how interactions are probed when relying on linear models. They proposed their "Binning estimator" and their "Kernel estimator" as alternatives. In this letter I identify a third problem that plagues interactions. It not only further invalidates interaction results from the linear models, but also invalidates results from these Kernel and Binning estimators. The problem arises when x and z in the *x·z* interaction are correlated and either has a non-linear effect on the dependent variable. I argue this third problem is likely to be ubiquitous in data used in political science in general, and show it's likely present in 8 out of 10 articles citing Hainmueller et al. I obtained as a convenience sample. I also show that GAM-based alternatives put forward in a recent article, "GAM Simple Slopes" and "GAM Johnson Neyman" (Simonsohn, in press) are not negatively impacted by this third problem, and enjoy other benefits such as greater precision, informativeness, scalability to additional covariates, and computational efficiency.

The supplementary materials, and the code to reproduce all results are available from:
https://researchbox.org/2063 (code **PLPFAX**)

Interactions are usually tested in social science by estimating a regression model like y=**a**+**b**x+**c**z+**d**x·z+ε. If $\hat{d}$ is statistically significant, authors typically conclude that z moderates the effect of x. This conclusion is often followed up by computing the marginal effect of x on y for different values of z, known in some disciplines as "probing" the interaction (Aiken & West, 1991; Preacher, Curran, & Bauer, 2006).

In an article with nearly 900 Google Scholar citations as of November 2023, Hainmueller, Mummolo, and Xu (2019) discuss two problems with probing interactions from linear models, "`First, these models assume a linear interaction effect that changes at a constant rate with the moderator. Second, estimates of the conditional effects of the independent variable can be misleading if there is a lack of common support of the moderator.`" (p.1; abstract).

They propose two alternative estimators for probing interactions: (i) the "binning estimator", which is a specific operationalization of a segmented regression, and (ii) the "kernel estimator", which is a specific operationalization of kernel regression more generally. Their article is excellent. It provides a clear exposition of the two problems that concerned the authors and demonstrates the use and interpretation of the two estimators they proposed, re-analyzing data from 22 published papers.

In this letter, however, I point out that Hainmueller et al. overlooked a *third* problem afflicting interactions. That problem, unfortunately, not only further invalidates the linear probing of interactions they critique, but it also invalidates the binning and kernel estimators they propose. The third problem, moreover, is ubiquitous in non-experimental data.

The problem is that if the two predictors in an *x·z* interaction are correlated, and either x or z has a non-linear effect on y, the linear regression estimate of the interaction effect, *x·z,* is biased, and therefore so are the marginal effects computed off such estimate. This was recognized long ago in psychology (Cortina, 1993; Ganzach, 1997, 1998; Lubinski & Humphreys, 1990), but it has been largely ignored by researchers and methodologists in psychology and beyond.

This problem is related but distinct from the one discussed in a recent letter by Beiser-McGrath and Beiser-McGrath (2023), which also focuses on the binning estimator by Hainmueller et al. The authors of the letter are concerned with misspecification of the impact of covariates outside of the *x·z interaction*. They propose adding polynomials of those covariates as controls (after a Lasso filtering process). That modification does not address the bias in the binning estimator discussed here, at all.[1]

In a recent article, I propose a solution to this problem of "*correlated non-linear predictors*" (Simonsohn, in press **| *available from*** [***http://urisohn.com/42***](http://urisohn.com/42)). After exploring 1000s of simulated scenarios, reanalyzing data from published papers, and considering several statistical solutions, I conclude that Generalized Additive Models (GAMs) provide the best tool for both testing and probing interactions (see Table 1 in that article).

In this letter I do not, and could not, repeat all the analyses from that paper. Instead, I focus on providing an intuition for why this neglected third problem invalidates the linear model and the proposed binning and kernel estimators, and why it does not invalidate the GAM-based solutions.

Bias from correlated non-linear predictors

Figure 1 below provides a stylized and concrete illustration. It considers a true model where x has a nonlinear effect, $y=x^2$, and where x is correlated with a third variable, z, which doesn't enter the true model (and thus doesn't moderate x). The figure shows that if we include z as a moderator in the statistical analyses, the linear model, and both of Hainmueller et al.'s estimators, lead to very similar, and very similarly invalid results. All three falsely conclude z moderates dy/dx. I explain the depicted GAM-based estimator in a later section.

[1]Their conclusions section reads, for example, that "there is the risk of unmodeled nonlinearities among variables **used for covariate adjustment** biasing interaction effect estimates" (emphasis added). The authors also propose estimating a Lasso with polynomials for all predictors to compliment the binning estimator. In Supplement 2 I show that GAM outperforms this approach: as Lasso produces mean-squared-errors between 170% and 350% as big as GAM does (when applied to the simulations I developed long before knowing I would evaluate how Lasso performs in them).

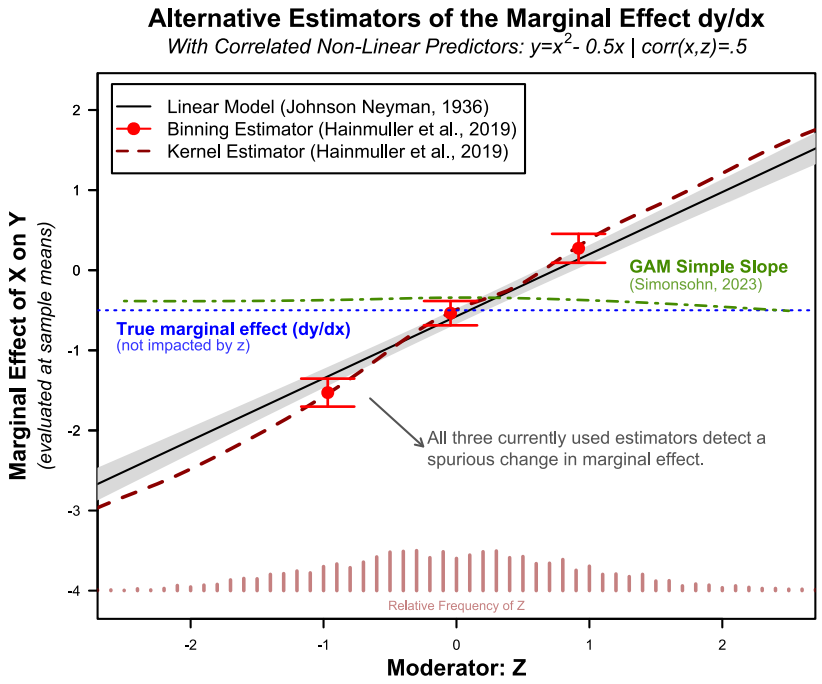

Fig 1. Correlated Non-Linear Predictors Produce Spurious Interactions
Results based on a single illustrative simulation with N=2500 observations. The binning and kernel results obtained with R package "interflex".
R Code to reproduce figure: https://ResearchBox.org/2063.7 (code: PLPFAX)

Figure 2 shows results for 1000 simulations like the one behind Figure 1, demonstrating the obtained results are not a fluke. The three existing estimators are systematically biased in the presence of correlated non-linear predictors.

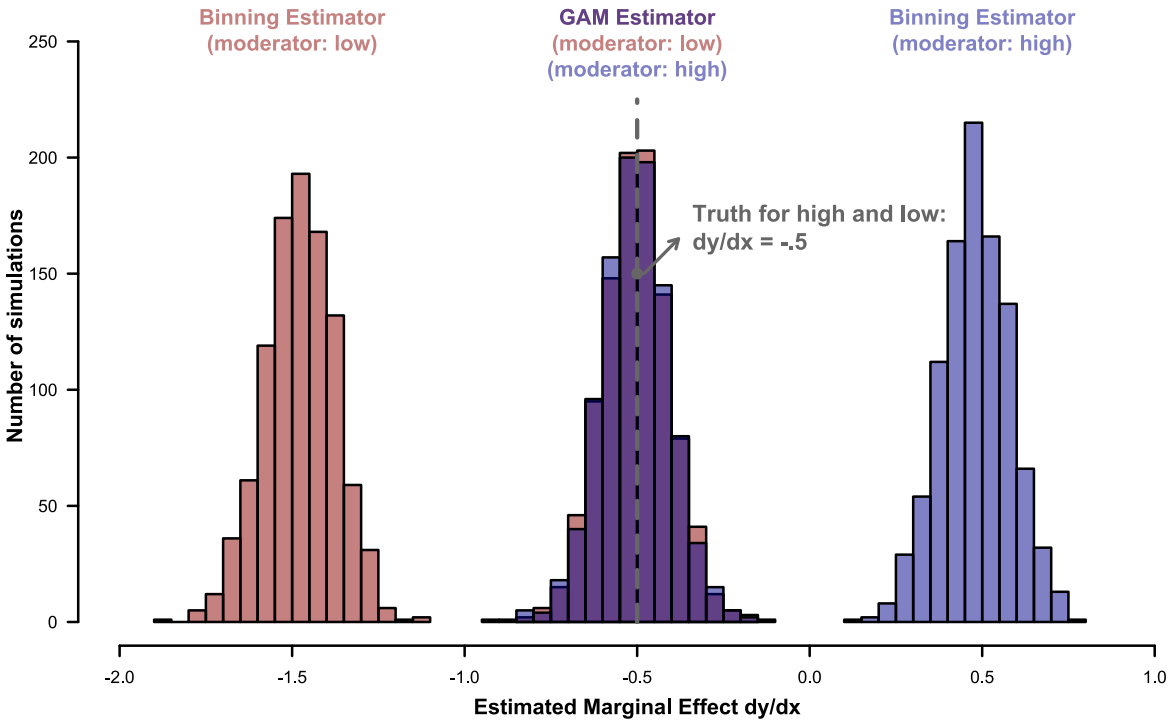

Fig 2. Marginal Effects with Binning vs GAM estimator
Histograms of estimated marginal effects across 1000 simulations of the data generating process behind Figure 1. The true marginal effect is dy/dx= -0.5 for all z (at sample mean of x=0).
R Code to reproduce figure: https://ResearchBox.org/2063.7 (code: PLPFAX)

Intuition for the problem and its solution

A useful framework for understanding the problem produced by correlated non-linear predictors is "omitted variable bias". It is well known that estimated coefficients in a regression are biased if the regression omits covariates that correlate both with the dependent variable and with predictors left in the regression. Quoting from my aforementioned article (Simonsohn, in press): "Studying interactions assuming the effects of *x* and *z* on *y* are linear, is equivalent to *omitting the non-linear portions* of the effects of x and z from the regression". Because of omitted variable bias, then, predictors in a regression that correlate with the omitted nonlinearities of x or z will have biased estimates. And, that is exactly what happens with the *x·z* interaction.

The quadratic example for Figure 1 makes the simile to omitted variable bias literal, as we are literally omitting the term x^2 from the regression; because z is correlated with x, it follows that $z \cdot x$ is correlated with $x \cdot x$, *the omitted* x^2 term. Thus $x \cdot z$ is biased by the omitted variable x^2.

It may seem surprising that the binning estimator is also biased by the correlated non-linear predictors problem, because it seems like it does not assume linearity. But, it does. The binning estimator is just a segmented *linear* regression; within each segment, the effects of x and z on y enter linearly as predictors. Omitted nonlinearities *within segments* produce bias within segments. Kernel regressions, in turn, do not *need* to impose linearity, but the kernel estimator proposed by Hainmueller et al. does, and thus is also biased.[2]

To test and probe interactions in a manner that is robust to correlated nonlinear predictors, we need estimators that flexibly estimate the functional form of the effects of x and z. A simple and efficient way to do this is through generalized additive models (GAMs). GAMs were developed decades ago (Hastie & Tibshirani, 1987; Wood, 2006), but have not been used much in social science. In a nutshell, GAMs *estimate* the functional form between predictors and the dependent variable, by combining a series of 'basis' functions when fitting the data. A penalty for excessive wiggliness in the resulting function curtails overfitting.

In the context of a model with two predictors with a possible interactions, a GAM can estimate the functional form of three additive functions: $y=f_1(x)+f_2(z)+f_3(x,z)$. For a brief and focused overview of GAMs for the purpose of studying interactions, see the subsection "Approach 3: Generalized Additive Models" in Simonsohn (in press). The textbook by Wood (2006) provides a thorough, rigorous, and general discussion of GAMs for a broad range of analytical situations. Section 7.7 in the textbook by James, Witten,

[2] The kernel estimator Hainmueller et al. propose is specified in their Equation 5. Using notation from this article, the key terms in that equation are $\mathbf{y=f(z)+g(z)\cdot x}$. The effect of z is allowed to be nonlinear, f() and g() are flexible, but the effect of x is not allowed to be nonlinear. Kernel regression can easily accommodate nonlinear effects in all predictors. For instance, that's the default behavior with R's np::npregres().

Hastie, and Tibshirani (2021) provides a brief introduction to GAMs, positioning it in relation to other approaches, both more and less flexible ones.

Hainmueller et al. do briefly mention GAMs (in footnote 1 and Appendix A.3), but they consider GAMs only for descriptive purposes, and only when both x and z in $x \cdot z$ are continuous. GAMs, however, are useful for *inferential* purposes, for *testing* whether the interaction between x and z is statistically significant. Indeed, I found that *only* GAMs provide valid tests of interactions in the presence of correlated nonlinear predictors (see Figure 11 in Simonsohn, in press). Moreover, GAMs *can* be used when x is dichotomous, and provide valuable details about functional form in those cases (see e.g. figures 4 & 5 in Simonsohn, in press).

A common concern with GAMs is that they produce estimates that are not easy to interpret (see e.g., James et al., 2021, p. 310). This concern arises because GAMs fit data by combining basis functions, and the default software output from a GAM estimation corresponds to the best-fitting weights given to those functions. That default output is indeed utterly uninterpretable by people.

A simple solution is to focus, instead of on the default output, on predicted values of the dependent variable, and estimated marginal effects, for different predictor values. Those results *are* interpretable. In the aforementioned article (Simonsohn, in press), I specifically proposed probing interactions from GAM models with "GAM Simple Slopes" and/or "GAM Johnson Neyman" curves.

The former involves plotting the focal predictor on one axis, x, and fitted y-values on the other, plotting separate lines for different values of the moderator, say the 85^{th}, 50^{th}, and 15^{th} percentile of z. I refer to these as "*GAM* Simple Slopes", because they generalize a procedure known as (linear) "Simple Slopes", a common approach for probing *linear* interactions (Aiken & West, 1991; Preacher et al., 2006). The only difference between linear and GAM simple slopes, from the perspective of the reader of a paper that reports them, is that the latter lines are not (necessarily) straight. GAM simple slopes are thus *identically* interpretable to those from linear regression.

The GAM Johnson Neyman procedure, in turn, generalizes the (linear) Johnson and Neyman (1936) procedure, which is similar to Simple Slopes, but with marginal effects, instead of fitted values, in the y-axis. Returning to Figure 1, the largely horizontal green dotted line, shows a GAM Johnson Neyman curve; the estimated dy/dx for all zs, when x is fixed at its mean of x=0.

This barebones R code produces a GAM Johnson Neyman line:

```
g1=gam(y~s(x)+s(z)+ti(x,z)) #Model with smooths for main & interaction effects
zs=seq(-2.5,2.5,.1)         #Set of z values to plot
yh1 = predict(g1,newdata = data.frame(x=.1,z=zs))  #Predicted when x=.1
yh0 = predict(g1,newdata = data.frame(x=0,z=zs))   #Predicted when x=0
gam.JN = (yh1 - yh0)/.1                            #dy/dx when x increases by 1
```

Is the problem of correlated predictors likely to matter in practice?

There are many methods papers that warn us of the terrible consequences of this or that assumption violation. Some of those papers warn us about scenarios that in practice almost never happen, others about scenarios that happen all the time. How often should we expect the problem of correlated non-linear predictors to matter? How worried should we be about the binning and kernel estimators proposed by Hainmueller et al. performing poorly in the real world? I think one should be quite worried.

For the problem to arise, two conditions need to be met: the true effects of x or z need to be non-linear, and x & z need to be correlated. Both things are likely ubiquitous in real non-experimental data. In terms of effects being nonlinear, there are good theoretical reasons to expect this. Psychology tells us that perception of physical and numerical stimuli exhibits diminishing rather than constant sensitivity. Economics tells us that marginal benefits and costs are decreasing and increasing respectively, not constant. Lastly, social science's general reliance on bounded scales to measure attitudes and beliefs mechanically produces nonlinearities through ceiling and floor effects. As Hainmueller et al. point out, linearity is a strong assumption.

In terms of the predictors in $x \cdot z$ being correlated. It seems likely that two variables hypothesized to impact the same dependent variable will be correlated. To get sense of the prevalence of $r(x,z) \neq 0$ for

x·z interactions that researchers typically test, I created a small but highly relevant sample: 10 highly cited empirical articles citing Hainmueller et al. (supplement 1 here has details). Only 2 of the 10 involved experiments, the other 8 studied *x·z* interactions where both x and z were measured rather than randomly assigned. None of the 8 articles discussed the problem of correlated nonlinear predictors, nor reported r(x,z). But r(x,z)=0 seemed unlikely in all 8 of them. I give two examples here, Supplement 1 covers all 10. The most cited article in the set is by Grossman et al. (2020). Their *x·z* interaction involves x: date when a governor tweeted asking people to stay home (during the Covid pandemic), and z: Trump's vote share in the governor's state. The paper indirectly documents that r(x,z)≠0; their Figure 1 shows that democratic governors (z), tweet earlier (x). In another article (Sands & de Kadt, 2020) the *x·z* interaction was composed of x: people's personal wealth and z:local Gini coefficients, these two wealth related variables are unlikely to be uncorrelated. To be clear: I am not proposing that if a published paper reports an *x·z* interaction and r(x,z)≠0 then we can conclude the results are wrong. It is not clear how many of the results in the literature would survive a GAM-based estimation, it is possible that all of them would (and possible that none of them would).

What about data from experiments?

The problem of correlated nonlinear predictors is absent in experiments; when x is randomly assigned, E(r(x,z))=0. Even here, however, GAM has some advantages. I discuss three. The first two are in relation to the binning estimator, the third in relation to both the binning and the kernel estimator. First, GAM-based estimates have slightly less sampling error than estimates from the binning estimator (see Supplement 3). The difference is quite small, but, given the greater flexibility of the GAM model, one might have expected the GAM to have *more* sampling error. Second, GAMs give much more information than the binning estimator does; functions are more informatively described through a full curve, than through three points chosen arbitrarily from within that curve. The third advantage is that GAMs can incorporate

covariates with flexible functional form. Neither the binning estimator nor the kernel estimator can do that.[3]

These advantages are real, but reasonable readers may, despite them, choose to analyze experimental data with Hainmueller et al.'s binning estimator. For them, a closing word of caution: *never use the binning estimator to test the interaction*.

Hainmueller et al. propose relying on the binning estimator to provide a "`formal test of the extent to which the data contains evidence of a significant interaction effect once we relax the stringent [linearity] assumption`" (p.21). Specifically, they propose testing the interaction by assessing whether the marginal effects in the top vs bottom bin are statistically significantly different. I think that's a bad idea. This binning estimator has substantially lower power than the linear regression, *and offers zero benefit* over the linear regression (for testing purposes).

In terms of power. A test comparing marginal effects at two (arbitrary and non-extreme) moderator values, has less power than one including *all* moderator values, which is what the regression coefficient for the interaction provides. For back-of-envelop calculations I simulated data where the true model was linear: when a regression had 80% power to detect an interaction, the binning estimator had about 60%.[4]

In terms of the lower-power binning estimator having no benefit. Recall that when x is randomly assigned, non-linearities in z, the moderator, do not correlate with the interaction and thus the omitted non-linearities of z do not bias the interaction coefficient, nor inflate its false-positive rate (Simonsohn, in press). There is therefore no need for an alternative interaction test for data from experiments. It is thus *never* beneficial to use the binning estimator for *testing* interactions. With non-experimental data, the

[3] While the kernel estimator put forward by Hainmueller et al. does not accommodate covariates with flexible functional form, kernel regression more generally does. In supplement 2 I contrast GAM with kernel regression in such situation.

[4] The simulation has x=0 or x=1, z~N(100,10) and $y=x+z+x\cdot z+\varepsilon$. With N=900, across 1000 simulations, 79.4% of interaction tests were significant, compared to just 57.8% for the top vs bottom bin. The p-value was larger for the latter test in 82.7% of simulations. R Code: https://researchbox.org/2063.17 (Code: **PLPFAX**)

binning estimator is an *invalid* test, with experimental data, the binning estimator is an unnecessarily *underpowered* test.

References.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*: Sage.

Beiser-McGrath, J., & Beiser-McGrath, L. F. (2023). The Consequences of Model Misspecification for the Estimation of Nonlinear Interaction Effects. *Political Analysis, 31*(2), 278-287.

Cortina, J. M. (1993). Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of management, 19*(4), 915-922.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological methods, 2*(3), 235.

Ganzach, Y. (1998). Nonlinearity, multicollinearity and the probability of type II error in detecting interaction. *Journal of management, 24*(5), 615-622.

Grossman, G., Kim, S., Rexer, J. M., & Thirumurthy, H. (2020). Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences, 117*(39), 24144-24153.

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis, 27*(2), 163-192.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association, 82*(398), 371-386.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs, 1*, 57-93.

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious" moderator effects": Illustrated substantively with the hypothesized (" synergistic") relation between spatial and mathematical ability. *Psychological bulletin, 107*(3), 385.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*(4), 437-448.

Sands, M. L., & de Kadt, D. (2020). Local exposure to inequality raises support of people of low wealth for taxing the wealthy. *Nature, 586*(7828), 257-261.

Simonsohn, U. (in press). Interacting With Curves: How to Validly Test and Probe Interactions in the Real (Non-linear) World. *Advances in Methods and Practices in Psychological Science*. doi: - https://urisohn.com/42

Wood, S. N. (2006). *Generalized additive models: an introduction with R* (1st ed.): Chapman & Hall/CRC Monographs on Statistics and Applied Probability.