Two-Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships with Quadratic Regressions

Uri Simonsohn[1]

[1]Affiliation not available

October 25, 2018

Abstract

Many psychological theories predict U-shaped relationships: the effect of x is positive for low values of x, but negative for high values, or vice-versa. Despite implying merely a change of sign, hypotheses about U-shapes are tested almost exclusively via quadratic regressions, imposing an arbitrary functional form assumption that can lead to a 100% false-positive rate, e.g., concluding with certainty that y=log(x) is U-shaped. Estimating two regression lines, one for low and one high values of x, allows testing for a sign change without a functional form assumption. To set the breakpoint between the lines, I introduce the Robin Hood algorithm. It delivers higher power to detect U-shapes than all other breakpoint setting alternatives considered. The paper includes simulations and re-analyses of published results. The two-line test can be performed at http://webstimate.org/twolines

This draft: July 11th, 2018

Two-Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships with Quadratic Regressions

Uri Simonsohn
Ramon Llull University
ESADE Business School
urisohn@gmail.com

Abstract:
Many psychological theories predict U-shaped relationships: the effect of x is positive for low values of x, but negative for high values, or vice-versa. Despite implying merely a change of sign, hypotheses about U-shapes are tested almost exclusively via quadratic regressions, imposing an arbitrary functional form assumption that can lead to a 100% false-positive rate, e.g., concluding with certainty that y=log(x) is U-shaped. Estimating two regression lines, one for low and one high values of x, allows testing for a sign change without a functional form assumption. To set the breakpoint between the lines, I introduce the Robin Hood algorithm. It delivers higher power to detect U-shapes than all other breakpoint setting alternatives considered. The paper includes simulations and re-analyses of published results. The two-line test can be performed at http://webstimate.org/twolines

Online supplement: https://osf.io/psfwz/

Is there such thing as too many options, virtues, or examples in an opening sentence? Researchers are often interested in such U-shaped relationships, where the effect of x on y is hypothesized to be positive for low-values of x, but negative for high values of x (or vice versa). Just among papers published online in 2016, for instance, I found at least two articles testing U-shaped relationships in each of the following four journals: JEP:G, *Psychological Science, JPSP,* and the *Journal of Applied Psychology*.[1]

Here I identify, and provide a remedy for, a large and pernicious disconnect between the predictions that social scientists make when they hypothesize that a relationship is U-shaped, and the statistical test social scientists run to examine if a relationship is U-shaped.

In particular, when a social scientist hypothesizes that a relationship, y=f(x), is U-shaped, she is merely hypothesizing that f(x) contains a sign change: for low values of x its effect on y is positive, f'(x)>0, while for high values of x the effect is negative, f'(x)<0. Or vice versa. As Lind and Mehlum (2010) write in their methodological paper on U-shape testing "to test . . . for the presence of a U shape . . . we need to test whether the relationship is decreasing at low values . . . and increasing at high values" (p.110).

Just a sign change is what the U-shape hypotheses from the eight papers in the opening paragraph predict, what all articles reviewed by Grant and Schwartz (2011) examining U-shaped effects in psychology hypothesize, what the classic Yerkes Dodson "law" involves, etc.

Nevertheless, when it comes to testing empirically whether f(x) is U-shaped, social scientists do not just examine if it exhibits a sign change. They instead estimate a quadratic

---

[1]The references for these articles are (i) Payne, Brown-Iannuzzi, and Loersch (2016), (ii) von Bastian, Souza, and Gade (2016), (iii) Choi and Kirkorian (2016), (iv) Loschelder, Friese, Schaerer, and Galinsky (2016); (v) Jaspers and Pieters (2016), (vi) Josef et al. (2016); (vii) Koopmann, Lanaj, Wang, Zhou, and Shi (2016), and (viii) Wilson, DeRue, Matta, Howe, and Conlon (2016)

regression of the form: $y=b\mathrm{x}+c\mathrm{x}^2$ and then rely on its estimates, which are valid only if the arbitrarily assumed quadratic functional form is exactly true, to valuate if the results imply a sign change in f(x).

We are so used to testing U-shapes with quadratic regressions that researchers often use the two terms as synonymous, but U-shapedness and quadratic describe distinct features of mathematical functions.[2] For instance, the relationship between the standard deviation and the variance is quadratic, $V=(SD)^2$, but it is not U-shaped. Conversely, y=log(x) - 2x, is U-shaped but not quadratic: the values of y are *not* proportional to the square of the values of x.[3]

Assuming a quadratic functional form, when the functional form is not quadratic, can elevate false-positive and false-negative rates of U-shapes. The former are especially likely when the true function, f(x), flattens out (e.g., a ceiling effect), because the quadratic is unable to generate a long plateau and so, when its functional form is forced on the data, the quadratic generates a spurious sign-change. For instance, the quadratic regression can, under realistic circumstances, conclude with near certainty that y=log(x) is a U-shaped relationship; a 100% false-positive rate (see Figure 2). It can also plausibly obtain 100% false-*negative* rate, concluding with certainty a blatant U-shaped relationship is not U-shaped (see Figure 3).

In this paper I propose we rely, instead, on two regressions lines to test for the presence of a U-shaped relationship, one for low values of x, the other for high values of x, verifying one slope is positive and the other negative. The advantage is that regression lines can diagnose the

---

[2] Haans, Pieters, and He (2016) provide a thorough and thoughtful review of the empirical literature testing U-shaped relationships in management, and yet they quite explicitly treat U-shaped and quadratic relationships as synonymous (see their abstract and footnote 1). The methodological paper by Miller, Stromeyer, and Schwieterman (2013) on testing interactions for curvilinear relationships distinguishes between merely curvilinear and U-shaped relationships, but for both they assume a quadratic function. Lind and Mehlum (2010)'s paper on U-shape testing does distinguish between U-shaped and quadratic functions, but all their demonstrations involve estimating quadratic regressions.

[3] The function f(x)=log(x)-2x is U-shaped because its slope, f'(x)=1/x-2, is positive for x<.5 and negative for x>.5.

sign of the average effect without making functional form assumptions about f(x). This two-lines approach has on occasion been used as an informal follow-up robustness test to the estimation of a quadratic regression (see e.g., Iribarren, Sharp, Burchfiel, Sun, & Dwyer, 1996; Qian, Khoury, Peng, & Qian, 2010; Seidman, 2012; Ungemach, Stewart, & Reimers, 2011).

The contributions of the paper are (i) explaining why we must discontinue relying on quadratic regression, in any way, to test hypotheses involving U-shaped relationships, (ii) formalizing the two-lines approach to testing U-shapes, and (iii) introducing the "Robin Hood" algorithm to identify the breakpoint for the two-lines, obtaining higher statistical power for U-shape detection than all alternatives considered.

Defining "U-shaped"

The symbol used to represent U-shaped relationships, the letter "U", is symmetric, consists of an uninterrupted line, includes a flat portion in the bottom, and includes both a negatively sloped and a positively sloped section. When social scientists refer to a relationship as "U-shaped" however, they imply only that last property: the sign change.

When data are not continuous, e.g., taking only 5 possible values, researchers and methodologists anyway use the "U-shape" label to describe an effect that flips sign (see e.g., Cohen, Cohen, West, & Aiken, 2003, p. 576; Simonton, 1976). When the function is not symmetric, e.g., exhibiting a negative effect for ages 15 to 75 years, and a positive one only for ages between 75 and 95 years, researchers use the "U-shape" label to describe the sign change as well (Jaspers & Pieters, 2016). When the functional form lacks a flat portion and the effect switches abruptly from negative to positive, researchers also use the "U-shaped" label to describe the sign change (Choi & Kirkorian, 2016, see their Figure 3). Relying on the same terminology,

in this paper I use the "U-shape" label to imply only a sign change in f(x), without implying any of the other characteristics of the letter "U" to f(x).

Neither the two-lines test proposed here, nor the quadratic regression based tests of U-shapedness, statistically distinguish between symmetric vs asymmetric U-shapes, continuous vs discontinuous U-shapes, or U-shapes with vs. without flat portions (the quadratic regression implicitly assumes f'(x) is continuous, but does not *test* whether it is). Thus, researchers interested in assessing these additional features of f(x) would require running additional statistical tests, not just U-shapedness test, whether they rely on the quadratic regression or on the two-lines test.

Two average slopes

In line with the previous subsection, let's formally define a function, y=f(x), as U-shaped, if there exists an x value, x_c, within the set of possible x values, such that the *average* effect of x on y is of opposite sign for $x \leq x_c$ vs $x \geq x_c$. The null hypothesis is that, H_o: No such x_c value exists, and the alternative hypothesis, H_A: At least one such x_c value exists.[4]

To test if the effect of x on y changes sign for $x \leq x_c$ vs $x \geq x_c$, we need to set the value of x_c, and then compute two average slopes, one for $x \leq x_c$ and one for $x > x_c$. I will discuss the issue of setting the breakpoint later on, for now let's focus on the benefits of using two regression lines to estimate the two average slopes.

---

[4] One could refine the definition to preclude more than one sign change (e.g., not classifying a W-shape as U-shaped), and could implement the testing by recursively applying the U-shape test to the two segments behind the U-shaped pattern. But, such refinement adds complexity and does not seem useful for the vast majority of cases where a U-shaped relationship is hypothesized; more than one sign change seems like a rather unusual prediction in the social sciences.

Linear regressions compute the average slope in the data for the effect of x on y, regardless of the underlying functional form (see e.g., Gelman & Park, 2008).[5] Therefore, to compute two average slopes, we may simply estimate two regression lines (one for $x \leq x_c$ and another for $x \geq x_c$). We can then reject the null of absence of a U-shape if the slopes are of opposite sign and are both statistically significant.

It is very important to understand that the regression estimate is the average slope for any functional form and thus we are *not* assuming the true function is linear when computing the average this way. To illustrate this point, say the true relationship is $y=x^2$, not linear, and the data consist of three observations, x=1,2,3, and thus y=1,4,9. A linear regression will correctly recover the average slope, which is 4 in this example. In particular, the slope between the first two points is (4-1)/(2-1)=3, between the last two (9-4)/(3-2)=5, and between the first and last (9-1)/(3-1)=4. So we have (3+5+4)/3=4. Similarly, if $y=x^4$, so that now y=1,16,81, $\hat{b}$ will be 40, the average slope among those three points.[6]

While the regression estimates will correspond to the average slope in the range of data no matter what underlying function form f(x) has, this does not mean that the two-lines test will be valid under all circumstances or that it constitutes a non-parametric test.

First, if the true relationship had more than one sign-change, for example, if it were W-, N-, or X-shaped, the two-lines test may correctly but misleadingly report that one portion has *on*

---

[5] In particular, regression estimates are the weighted average of the slope of every pair of data points, weighting each pair by the square of the distance between predictor values. For instance, in the simple case with one predictor: $\hat{b} = \sum_{i,j} \frac{y_i - y_j}{x_i - x_j} (x_i - x_j)^2 / \sum_{i,j} (x_i - x_j)^2$. See e.g., Gelman and Park (2008).

[6] You can verify running this code in R:
```
x=1:3
y=x^4
lm(y~x)
```

average a positive slope and the other a negative one, leading a researcher to erroneously classify a W-, N- or X- shaped relationship as U-shaped (for more on this, see limitations section).

Second, because the two-lines test relies on linear regression, anything that affects the validity, interpretability, bias, robustness, efficiency, etc., of linear regressions, also affects the validity, interpretability, bias, robustness, efficiency, etc. of the two-lines test. For example, lack of independence across observations leads to under-estimated standard errors in regression results in general, and to higher false-positive U-shape results with the two-lines test in particular.

The misuse of quadratic regressions to test for U-shapedness

The sophistication with which results from quadratic regressions are interpreted, for U-shape testing, can be classified into three levels; they differ in how many additional calculations are conducted upon estimation.

Level 1: is the quadratic term significant?

The most basic approach involves checking if the estimates of *a* and *b*, in $y = ax + bx^2$, imply a U-shaped function and if b's estimate is statistically significant. This approach is advocated in some prominent textbooks. For example, Cohen, Cohen, West, and Aiken (2003) write, "The [quadratic] coefficient is negative [and significant]. . ., reflect[ing] the hypothesizes initial rise *followed by decline*" (p.198; emphasis added). The significant coefficient need not, in fact, imply a U-shape relationship.[7]

---

[7] In a later section Cohen et al. do warn against blindly relying on the quadratic terms writing "it is always important to examine the actual data against both the polynomial regression and some nonparametric curve such as lowess" [lowess: locally weighted scatter-plot smoothing]. (p.207). I do not, moreover, believe the authors would have fallen prey to such fallacious conclusions, but many readers of the text probably have.

A JPSP paper by Simonton (1976), with about 150 citations, illustrates. He established correlates of the eminence of 'geniuses'. One key inference was that "*ranked eminence [is a] curvilinear inverted-U function of education*" (abstract, *p.* 218). The point estimates of interest, within a larger specification, were $y=4.872 x - 11.96 x^2$ where *y* is the measure of eminence, and *x* of education (see estimates in his Table 2; p. 223).
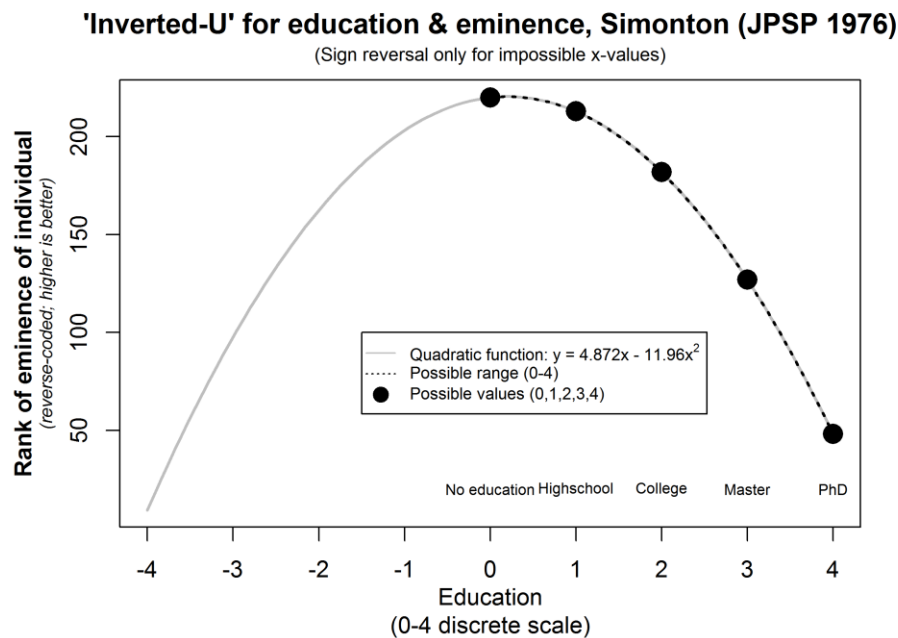

Fig. 1. *Example of significant quadratic term not associated with actual* U-shape.
R Code to reproduce figure: https://osf.io/9uwxg/

Figure 1 shows that, within the range of *possible* values, the regression results do not imply a U-shape. For every possible value of x, higher x is associated with lower y. Only for negative (impossible) values of x is the sign positive and hence the overall pattern U-shaped. Note that the estimated correlation between education and eminence is opposite of the intuitive causal effect one may expect.

Level 2: is the sign-flip within the range of values?

The discussion of Figure 1 above is an example of this additional step and some published papers carry it out as well. For example, Berman, Down, and Hill (2002) write "*the value [at which the sign flips] is actually above any value observed in the data, suggesting that, although negative returns are a theoretical possibility, they are not encountered.*" (*p.* 23).

But, we need to take into account sampling error. The true relationship is, we assume, $y=a\mathrm{x}+b\mathrm{x}^2$, but we don't observe *a* and *b*, we observe estimates, $\hat{a}$ and $\hat{b}$, and as estimates they contain error, and so does, therefore, our estimate of the point at which the effect of x on y flips sign ($-\hat{a}/2\hat{b}$).[8]

Level 3: is the sign-flip "statistically-significantly" within the range of values?

Noting that a quadratic term is simply an interaction of a variable with itself (see e.g., McClelland & Judd, 1993, p. 382), we can take into account sampling error into the analyses of quadratic regression estimates in general, and of the point where the effect of x on y flips sign in particular, as we do with any regression interaction. In particular, we may estimate the effect of x on y, and its confidence interval, or *p*-value, for different values of x. This general approach to analyzing interactions was first introduced by Johnson and Neyman (1936). The approach is sometimes known as pick-a-point or spotlight when applied to a handful of x-values, and as floodlight or Johnson-Neyman procedure, when applied to *all* of them, or to the critical x values where the slope is vs is not statistically significant (Aiken & West, 1991; Preacher, Curran, & Bauer, 2006; Spiller, Fitzsimons, Lynch Jr, & McClelland, 2013).

---

[8] If f(x) =ax+bx^2, then f'(x) = a+2bx. Solving for f'(x$_c$)=0 leads to x$_c$= -a/2b.

In recent years, a few papers have explicitly suggested relying on this Johnson-Neyman procedure to analyze quadratic regressions results for testing U-shaped relationships (Lind & Mehlum, 2010; Miller et al., 2013; Spiller et al., 2013). [9]

Importantly, even this more sophisticated use of quadratic regressions to test for U-shapedness, is invalid. The reason is that the regression results hinge, and therefore so do the Johnson-Neyman calculations, on the assumption that the true relationship between x and y is exactly quadratic. Figure 2 provides realistic examples where the assumption is not met, and the conclusions are erroneous.

For instance, panel A shows a scenario where the true relationship is y=log(x) and where a quadratic regression would result in $\hat{y}$ =13.96x-10.45x^2. In this equation, the effect of x on y is, dy/dx=13.96- 2*10.45x. When x=.25, the effect of x is positive, 8.735, but when x=.75, in contrast, negative, -1.71. Now, of course that result is wrong, the effect of x is never negative when y=log(x), but it is *estimated* as negative because we are incorrectly assuming the relationship is quadratic. Specification error is behind the erroneous conclusion. Panels B & C provide additional examples.

---

[9] Lind and Mehlum (2010) accompanied their (economics) paper with a STATA module, *utest*, that runs their proposed U-shape test. The program is executed after running a regression. When run after a quadratic regression, as all the examples in their paper do, their test is equivalent to the analysis advocated for in psychology textbooks, e.g., Aiken and West (1991, p. 77), see Supplement 4 for numerical demonstration of the equivalence. But Lind & Mehlum appear to have developed their test independently. The procedures by Spiller et al. (2013) and Miller et al. (2013) are, as these authors make clear, also directly derivable from the formulas in Aiken and West (1991, p. 77).
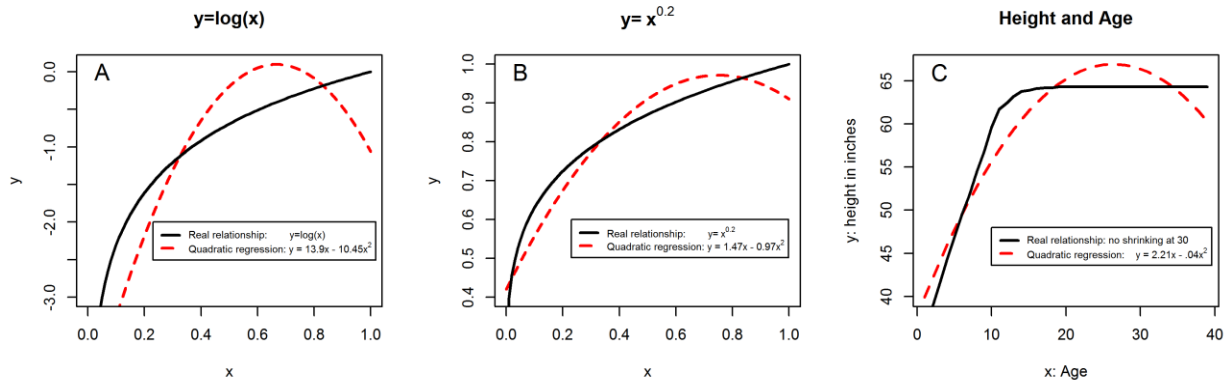
Fig. 2. *Examples of quadratic regressions misdiagnosing U-shapes.*
Panels **A** & **B** are obtained from a single simulated dataset with N=100000 where *x* is obtaining by squaring random draws from the U(0,1) distribution. Large samples without noise were used to convey the point that quadratic regressions get it wrong even asymptotically. Panel **C** uses data come from the Center for Diease Control.
R Code to reproduce figure https://osf.io/3psev/

Assuming a quadratic relationship may also lead to false-negatives, failing to diagnose a

U-shape relationship that is present, even with an *infinite* sample size. This will occur when the

true relationship is U-shaped but deviates sufficiently from the quadratic. See Figure 3.

The intuition for the poor performance of the quadratic regression in Figures 2 & 3 is that

it minimizes the sum of squared errors, $(\hat{y} - y)^2$, without taking into account overall-shape. If

obtaining a better fit requires outputting a quadratic function that generates a non-existent U-
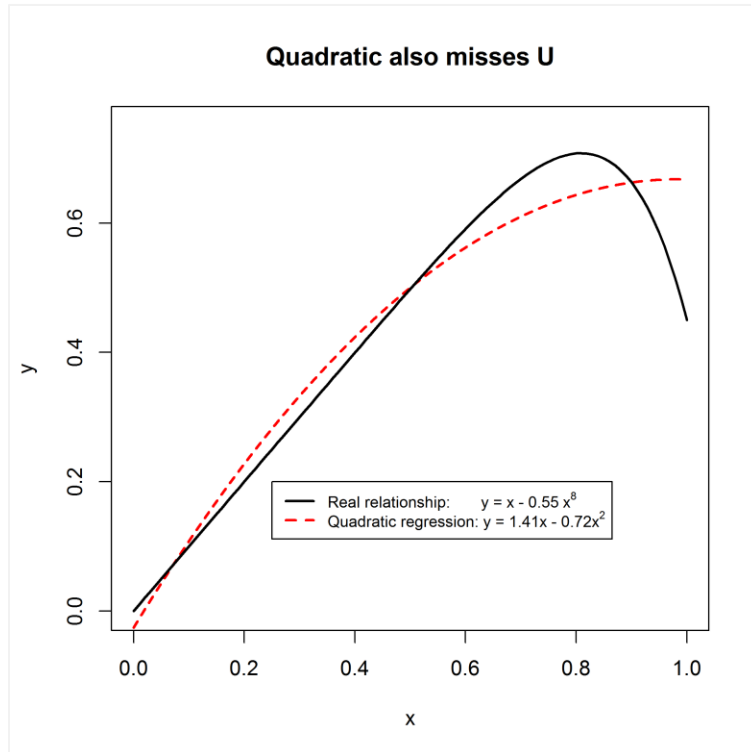
shape, or missing a real U-shape, there is no 'penalty.'

Fig. 3. *Example of quadratic regression false-negatively concluding U-shape is absent.*
A single dataset with N=100000 observations was created, x was generated by drawing at random from U(0,1) distribution and squaring the result. R Code to reproduce figure https://osf.io/3psev/

What about diagnostic tests?

Many textbooks indicate researchers should conduct diagnostic tests before interpreting their regression results, are those enough to protect us from wrong inferences about U-shapes based on quadratic regressions? I argue below the answer is no.

First, in practice, researchers do not run or at least report diagnostics on their regressions. Second, regression diagnostics qualitatively assess the general adequacy of the model, but we want to quantitatively assess the adequacy of the conclusion *about U-shapedness*. Figure 4 illustrates the problem, where regression diagnostics for true-positive and a false-positive U-shaped relationships are indistinguishable from one another.

Third, it is not clear what researchers should do when they diagnose their quadratic regression as misspecified. If not a quadratic model, then what model should they estimate? There is no default alternative, researchers would need to try multiple functional forms until one seems to -subjectively- fit well enough. Say running higher order polynomials, interrupted log regressions, various interactions, etc. This leads to two problems. First, when those more complicated models are estimated, how would the researcher go about testing for U-shapedness?
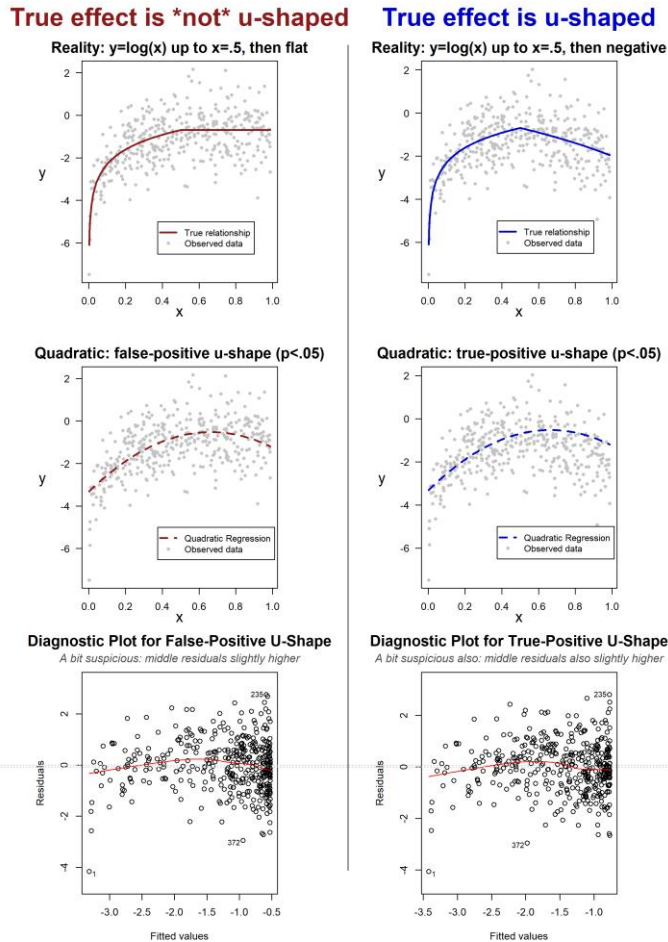
True effect is *not* u-shaped   **True effect is u-shaped**

Fig. 4. *Diagnostic plots are not diagnostic about U-shapedness*
Notes: The data were generated by drawing 400 observations from U(0,1) for x and adding noise N(0,1) to the true y-value (see top-row for true model). Each column has the same dataset for the three charts, but they differ across columns.
R Code to reproduce figure: https://osf.io/kuj3d/

For example, if we fit a fourth order polynomial to the y=log(x) data from Figure 2A, we obtain the following estimate: $y = 44\,x - 142\,x^2 + 189\,x^3 - 86\,x^4$. Should we interpret this equation as evidence for or against a U-shape? Perhaps the most sensible thing to do is to compute the implied marginal effect of x on y for every value of x, and then average them for two ranges of x. But now we have a two-lines test, except that we are averaging *fitted* data, through an arbitrary functional form, instead of raw data.

14

In addition to the inherent ambiguity of results from higher order polynomials or other arbitrary functional form assumptions, and likely invalidity of such assumption, the mere fact that there are so many alternative procedures opens the door to over-fitting in general and *p*-hacking in particular.

The two-line solution

Because U-shaped hypotheses state merely that the effect of x on y changes sign for low vs high x values, we should test U-shaped hypotheses by merely testing if the effect of x on y changes sign for low vs high x values. This involves computing two average slopes, which in turn is done by estimating two regression lines, one for $x \leq x_c$ the other for $x \geq x_c$, where x_c is the breakpoint separating the two regions. One may increase statistical efficiency by simultaneously estimating both lines in a single regression, relying on what is often referred to as an *interrupted* regression (see e.g., Marsh & Cormier, 2001, p. 7). Specifically, interrupted regressions conform to the following general formulation: [10]

$$y = a + b \; x_{low} + c \; x_{high} + d \; high + Z \; B_z \qquad (1)$$

Where: $x_{low} = x - x_c$ if $x \leq x_c$, and 0 otherwise

$x_{high} = x - x_c$ if $x \geq x_c$, and 0 otherwise

high = 1 if $x \geq x_c$, and 0 otherwise.

Z is the (optional) matrix with covariates, and B_z its vector of coefficients.

---

[10] If d is forced to be 0, thus not allowing a discontinuity at x_c, the regression is called *segmented* instead of *interrupted* (see e.g., Muggeo, 2003). Forcing d=0 introduces bias onto both $\hat{b}$ and $\hat{c}$. For U-shape testing purposes, one must rely on interrupted rather than segmented regressions, include *high* as a predictor.

Setting the breakpoint

We can set the breakpoint seeking to maximize fit, *or,* seeking to maximize statistical power. That is, seeking to arrive at a model that fits the data best, *or* at one that has the highest probability of diagnosing f(x) as U-shaped when it is, without exceeding the nominal false-positive rate when it is not.

Maximizing fit. Setting the breakpoint to maximize fit involves answering this question: "Given that we will fit the data with two lines, which breakpoint leads to two lines that best fit the data overall?" There is a literature examining how to set breakpoints, in segmented and interrupted regressions, seeking to fit the data as well as possible, e.g., via maximum likelihood (see e.g., Hansen, 2000; Molinari, Daures, & Durand, 2001; Muggeo, 2003; Stasinopoulos & Rigby, 1992). But that is not what we want to do here.

We are not fitting two lines, with a possible discontinuity between them, because we believe the real relationship has that shape and we want to approximate it as well as possible. Rather, we are only estimating regressions to compute average slopes in two sets of x-values.

Thus, we want to find the breakpoint that answers a different question: "If the true relationship were U-shaped, if there really were a sign-change for the effect of x on y within the set of observed values, which breakpoint maximizes the chance we will detect it?" Figure 5 illustrates the conflict between these two goals. Moreover, later on, when evaluating the performance of different breakpoints, I will show that the breakpoint that maximizes fit (in that case R^2), obtains lower statistical power than that obtained with the proposed Robin Hood procedure.
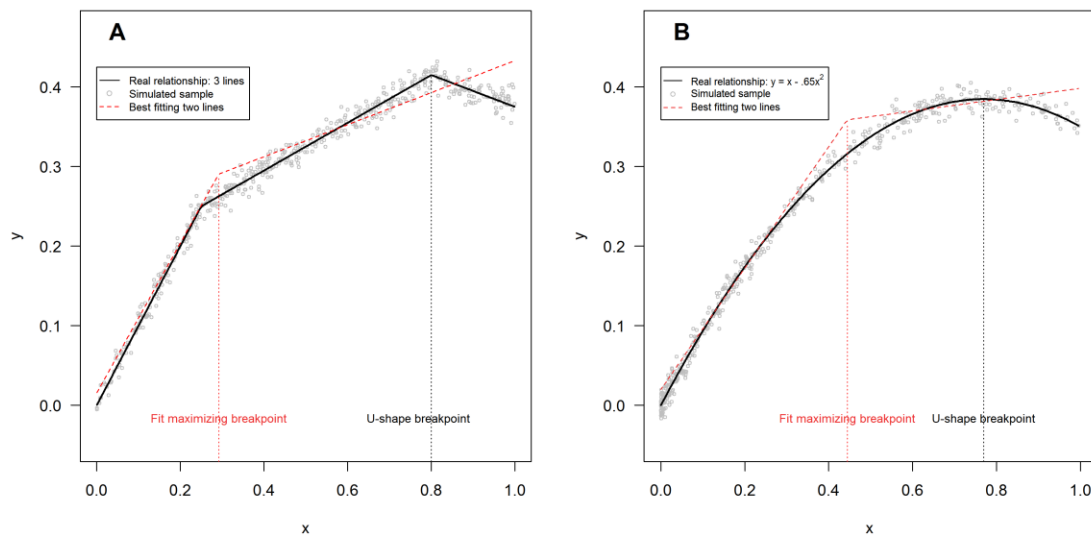
Fig 5. *The breakpoint that maximize overall fit does not necessarily maximize power to detect a U-shaped relationship.*

Note: the figure depicts the breakpoint for two regression lines that maximizes overall fit, using Muggeo (2003)'s procedure, and contrasts it with the breakpoint actually associated with the x-value at which the sign of the effect of x on y changes. R Code to reproduce figure https://osf.io/w3m2u/

Maximizing power. Without making strong assumptions about (a) the functional form of the relationship between x and y, f(x), (b) the distribution of x, and (c) the distribution of the error term, it does not seem possible to arrive at a theoretically optimal breakpoint that maximizes statistical power for U-shape testing. The approach I propose here, instead, is algorithmic, designed to have high power, rather than demonstrably maximal power, for a very broad range of situations (but presumably not all). I developed the algorithm keeping in mind three key ideas: (i) because the two-lines test requires both slopes to be significant, to increase its power requires increasing the power of the statistically weaker of the two lines. Segments of an interrupted regression, in turn, have more power when (ii) they are steeper (bigger effect), and (iii) they include more observations within their segment (smaller standard error). Thus, conceptually, the algorithm seeks to set a breakpoint that will increase the statistical strength of the weaker of the two lines, by placing more observations in that segment, without overly

17

attenuating its slope. I refer to it as the Robin Hood algorithm, for it takes away observations

from the more powerful line and assigns them to the less powerful one.

I rely on Figure 6 to describe the Robin Hood algorithm. Every panel involves the same

true underlying relationship between x and y, depicted by the solid line in Panel A, and the same

single random sample, depicted with the same scatterplot in every panel. The top row illustrates

increasingly more sophisticated approaches for setting the breakpoint, culminating in the

proposed Robin Hood algorithm in the right-most column. The bottom row the resulting two-line

regression estimates.

For illustrative purposes, consider attempting to obtain two steep slopes by setting x_c, the

breakpoint, at the x value associated with the most extreme observed y value (first column in

Figure 6). An obvious problem is that individual observations, especially the most extreme one,

can be greatly influenced by random error. Panel A, for example shows that the x value

associated with the most extreme observation, x=.78, falls outside the range with maximum true

y values, .5<x<.7.

We can cancel much of the aforementioned random error by estimating a flexible model

of f(x), e.g., a polynomial, local, kernel, or spline, regression, and use the model's fitted values

instead of the observed values to identify the most extreme observations.

I rely on splines here, because they easily accommodate covariates, can be used to

construct confidence intervals for f(x), and do not rely on functional form assumptions (see

section 3.2.1 in Wood, 2006).[11] In particular, Panel B depicts the fitted values, $\hat{y}$s, obtained from

a cubic spline regression, and showcases the consequences of moving the breakpoint from the x

---

[11] In particular, using the R library *mgcv*, the command gam(y~s(x,bs="cr")) estimates a cubic spline predicting the dependent variable y with the predictor x. The option bs="cr" specifies a cubic spline be used, instead of the default which is a "plate regression spline" (Wood, 2006, p. 219).

associated with the most extreme observed y, to the x associated with the most extreme *fitted*

value: $\hat{y}_{max}$.

      In the example from Figure 6, and presumably in many psychological phenomena,

relationships are U rather than V shaped, having regions with a relatively flat maximum. It seems

therefore sensible to identify the *set* of most extreme $\hat{y}$s rather than the single most extreme $\hat{y}_{max}$.

Here I define $\hat{y}$s within one standard error of $\hat{y}_{max}$ as that set and refer to it as $\hat{y}_{flat}$. Thus, every

$\hat{y}$ in $\hat{y}_{flat}$ is within one standard error of $\hat{y}_{max}$. The solid line in Figure 6C depicts $\hat{y}_{flat}$.

      We now have a set of candidate x_c values, those associated with $\hat{y}_{flat}$. The goal is to

choose the one among them that we expect to give higher statistical power to detect a U-shape,

and thus the one among them that we expect to give higher statistical power to the weaker of the

two lines within the interrupted regression. The algorithm pursues that goal by setting x_c so that

it allocates a disproportionate share of the observations in $\hat{y}_{flat}$ to the weaker line; by increasing

the number of observations in that segment, it reduces its standard error, increasing its statistical

power.

      The algorithm proceeds in two steps. In the first step it identifies which of the two lines is

statistically weaker. In the second step it sets the breakpoint by allocating observations in $\hat{y}_{flat}$ to

the first vs second line in inverse proportion to their relative statistical strength. Specifically, in

the first step the algorithm sets the x-value that is the midpoint of $\hat{y}_{flat}$ as an interim breakpoint. It

estimates an interrupted regression and computes the (absolute value of the) test-statistics for

both lines, t_1 and t_2, and then sets the breakpoint for the second step in inverse proportion to these

ts. That is, the breakpoint becomes the $t_2/(t_1+t_2)$ percentile of the x-values within $\hat{y}_{flat}$.
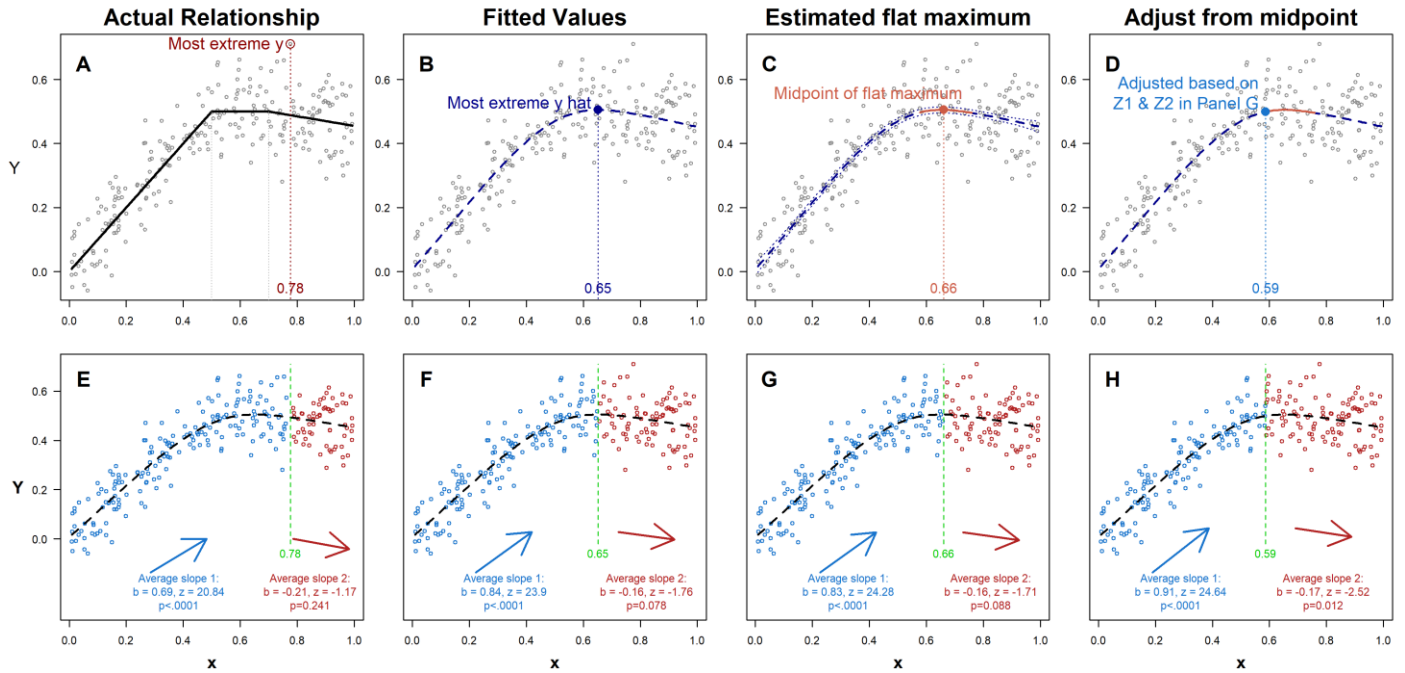
Fig 6. *Different procedures to identify the breakpoint, and their consequences.*
Notes: All panels are based on the same random sample (gray scatterplots) based on the true relationship between x and y, solid line in Panel A. The effect of x on y is positive up to x=.5, flat up to x=.7, and negative onwards. Top row shows 4 alternative ways to set the breakpoint, bottom row the resulting two-line regressions. Fitted values in panels B-D obtained by smoothing the scatterplot with a cubic spline. Flat region in C&D is where $\hat{y}$s are within 1 standard error of the max($\hat{y}$). R Code to reproduce figure: https://osf.io/zdert/

To build an intuition: if both lines are about equally strong, statistically speaking, with roughly identical test statistics, the breakpoint will remain roughly at the midpoint of $\hat{y}_{flat}$. If the t-value of the first line in the first step were, say, 3 times that of the second line, then the breakpoint would be set at the 75th percentile of xs within $\hat{y}_{flat}$, so that the second (weaker) line has 75% of $\hat{y}_{flat}$ and the first line the remaining 25%. The intuition, again, is that the algorithm allocates additional observations from within the $\hat{y}_{flat}$ region to the weaker line so that its standard error gets smaller.

Returning to Figure 6. Panel G shows that first step, where the midpoint of $\hat{y}_{flat}$ is the breakpoint. It leads to t_1= 25.07 and t_2=1.86. Computing the ratio we obtain $t_2/(t_1+t_2)$=6.9% so the breakpoint is set at the 6.9th percentile of the x values associated with $\hat{y}_{flat}$, which in that

sample corresponds to x=.59. Using that breakpoint, we obtain the final interrupted regression used to test the presence of a U-shape, and in this case, we obtain a much stronger result for the second slope, p_2=.006.

In sum, the Robin Hood algorithm consists of the following 5 steps.

1) Estimate a cubic spline for the relationship between x and y

2) Identify $\hat{y}_{max}$, the most extreme internal fitted value.

3) Identify $\hat{y}_{flat,}$ the set of $\hat{y}$ values within a standard error of $\hat{y}_{max}$

4) Estimate an interrupted regression using as the breakpoint the median x value within $\hat{y}_{flat}$. The regression will result in two test statistics, one for each line. Let their absolute values be t_1 and t_2

5) Set as the breakpoint at $t_2/(t_1+t_2)$ percentile of the x values associated with $\hat{y}_{flat.}$

It is important to note that because the breakpoint is set algorithmically within a set of candidate breakpoints, it conveys no interpretable meaning on its own. We should not conclude that the breakpoint is "the" point where the sign of the effect switches. The specific sign switch point, to the extent it actually exists, it not estimated precisely with the two-lines test.

Performance of two-line test

False-positive and False-Negative U-Shapes

I conducted simulations for a broad range of scenarios (see notes for Figures 8 & 9 for details) to assess type-1 and type-2 errors for examining U-shapes with a quadratic regression vs. the two-line test. For the two-lines approach I considered not just the breakpoint identified by the Robin Hood algorithm, but also for various alternative approaches, including maximizing fit.

21

For the quadratic regression approach I report results for the most sophisticated use of it, the procedure proposed by Lind and Mehlum (2010), which is equivalent to that proposed by Spiller et al (2013), and by Miller et al (2013), and by Aiken & West (1991).[12]

Figure 7 reports false-positive estimates. The left panel includes simulations where the true relationship is expected to lead to the most false-positive U-shapes: an initial strong effect, followed by a long flat segment.   The right panel for a (monotonic) logistic function.

The results from both panels are highly consistent. They show that the quadratic regression approach to testing U-shaped relationships has an unacceptably high false-positive rate for a very broad range of scenarios. Often 100%. The two-line approach in general, and the Robin Hood procedure for setting the breakpoint in particular, show, in contrast, acceptable performance. False positive rates are typically below the 5% nominal level (as is typically the case when the null hypothesis is a composite null (see e.g., Bowman, Jones, & Gijbels, 1998)) and even the post-hoc most extreme scenario raises it only barely above it (and these are necessarily over-estimates as we are selecting them ex-post *because* they were the highest value).

In Supplement 1 (https://osf.io/6c5qb/) I explore factors that lead to the higher false-positive rates, finding that scenarios that use the distribution of x suggested by McClelland (1997), see Figure 7 captions for a description, and those with greater levels of random noise, have higher false-positive rates. I run additional simulations that rely on that distribution of x and have even higher levels of noise, finding that the false-positive rate does not increase any further.

Figure 8 reports statistical power estimates for two general functional forms (Figure 9 shows a few examples of the individual simulations behind the left panel of Figure 8). Because the quadratic regression obtains unacceptably high false-positive rates, Figure 8 does not include

---

[12] Lind & Mehlum consider other functional forms in the theory section of their paper, but all their examples involve quadratic regressions.

power results for it. For statistical inference, we should select the most powerful test, *among those that satisfy the nominal false-positive rate.* To make this concrete, imagine a test consisting of a coin that reads "U-shape" on either side: flipping the coin leads to 100% power, but it is not a statistical test we would want to use.
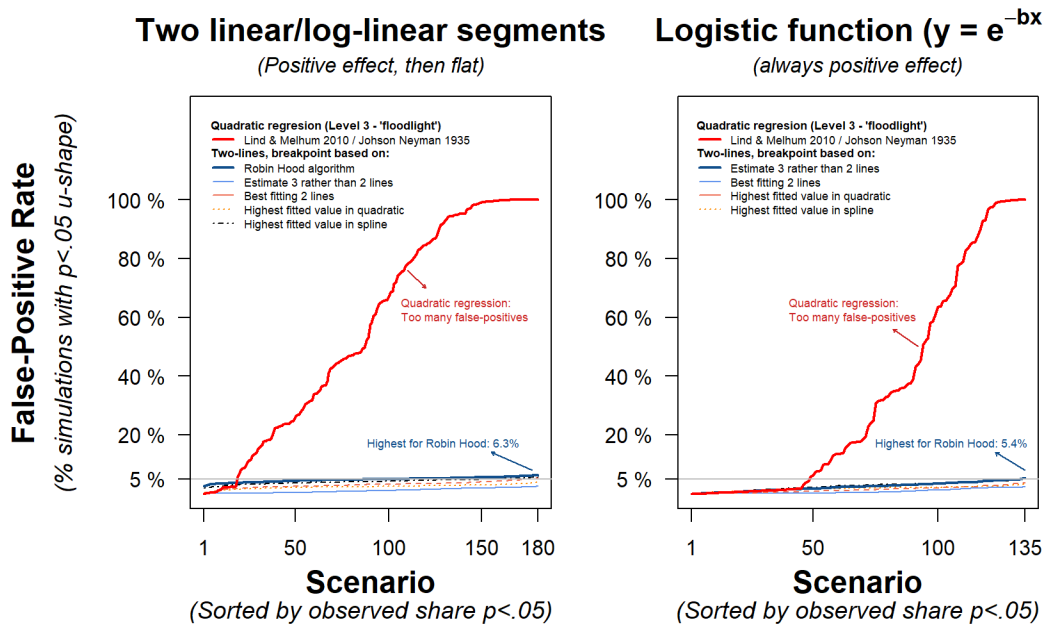

Fig. 7. *False-Positive Rates for Detecting U-shapes*
Notes: In the left panel each simulated scenario involves a relationship between x and y consisting of two segments. For $x<x_c$, the marginal effect of x on y is positive, for $x \geq x_c$ it is zero. The scenarios combine the following parametrizations: (i) the distributions of x (normal, uniform, beta with left, beta with right skew, optimized for the quadratic as in McClelland (1997)) [13], (ii) the effect of x on y is y=x vs y=log(x), that is, linear vs log-linear, (iii) sample sizes of 100, 200, or 500, (iv) σ in $e \sim N(0,σ)$ with σ being 100%, 200% or 300% of the SD(y) before adding *e*, and (v) the value of x_c: 30th, 50th percentile of x. The full combination of parameters leads to 180 scenarios. For the right panel the same parametrizations for sample sizes, and amount of noise are used. Instead of defining slopes and cutoffs, the scenarios differ in the values of b in the logistic function, with b=.5, 1.5, and 2.5, for a total of 135 scenarios. All alternative two-line estimations are for *interrupted* regressions (allowing a discontinuity at the breakpoint).[14]
R Code to reproduce simulations: https://osf.io/wdbmr/, figure itself: https://osf.io/q4ysc/

---

[13] In particular, 25% of observations are x<.2, 25% x>.8, and 50% are .4<x<.6 (McClelland, 1997; Table 1). This distribution is said to maximize power to detect a U-shape if the true relationship is quadratic and the maximum value is obtained at "intermediate values of X" (p.9).
[14] For each scenario 1000 simulations were run. If the false-positive rate was >4% for Robin Hood, an additional 10,000 simulations were run.

To facilitate comparisons with the proposed Robin Hood procedure, the share of statistically significant results for each procedure are presented, in Figure 8, as the difference with the Robin Hood's procedure. Both panels of Figure 8 also paint a highly consistent picture. For the vast majority of scenarios considered, Robin Hood provides a moderate to large improvement in statistical power, and for only a small minority it is not the single most powerful procedure, and even in those cases, the power losses are minimal. In addition to the superior performance of the Robin Hood procedure, two general patterns are worth highlighting. First, estimating three rather than two lines leads to dramatic losses of statistical power; the intuition is that the observations allocated to the middle line do not contribute to the precision of the slope estimates involved in the testing of the hypothesis. Second, the least powerful approach to setting the breakpoint for a two-line estimation, consists of the approach that has been previously proposed by several authors, including me, setting the breakpoint based on the quadratic regression's most extreme fitted value (Haans et al., 2016; Iribarren et al., 1996; Simonsohn & Nelson, 2014).
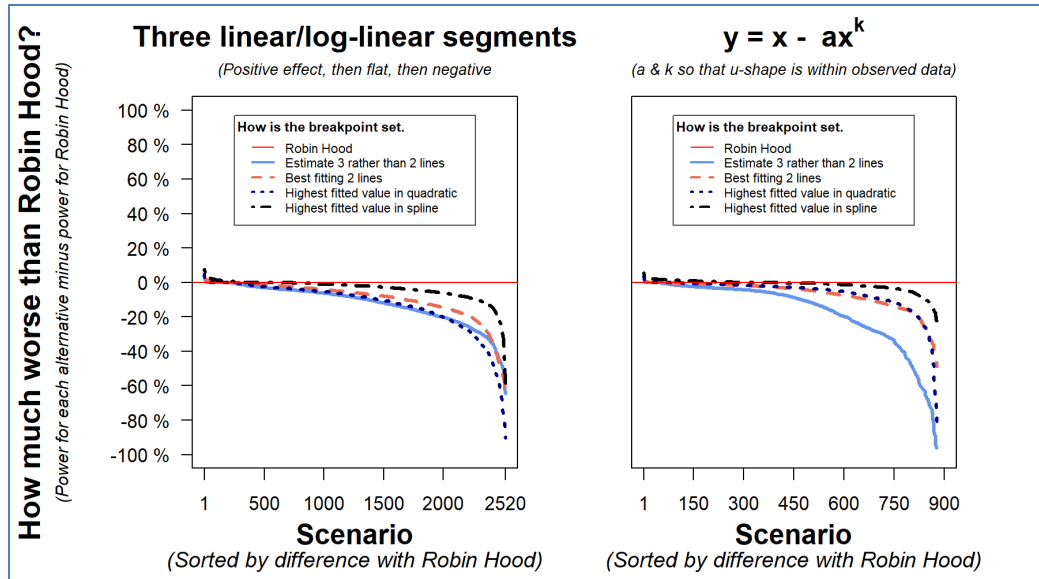
Fig. 8 *Statistical Power for Detecting U-shapes, relative to Robin Hood*

Notes: In the left panel each simulated scenario involves three segments, with cutoffs at x_c and x_d, with $x_d > x_c$ For $x < x_c$, the marginal effect of x on y is positive. When $x_c \le x \le x_d$ the marginal effect of x is zero, and when $x > x_d$ the marginal effect is negative. The scenarios combine the same parameters as Figure 8A, crossed with the value of x_d : 30th, 50th, 70th, or 90th percentile of x, and the slope of the negative effect of x on y when $x > x_d$ being 25%, 50%, 100% or 200% the magnitude of the slope when $x < x_c$. The full combination of parameters leads to 2520 scenarios (See Figure 9 for a few examples). For the right panel the same parametrizations for sample sizes, and amount of noise are used. Instead of defining slopes and cutoffs, the scenarios differ in the values of k, with $k=2,3,4$, or 5, and the values of *a,* which are set so that the resulting U-shape provides a sign change at 50%th, 60th, 70th, 80th or 90th quantile of x of observed values. Each of the 3420 scenario results are based on 500 or 2500 simulations, depending on the extremity of results after 500.

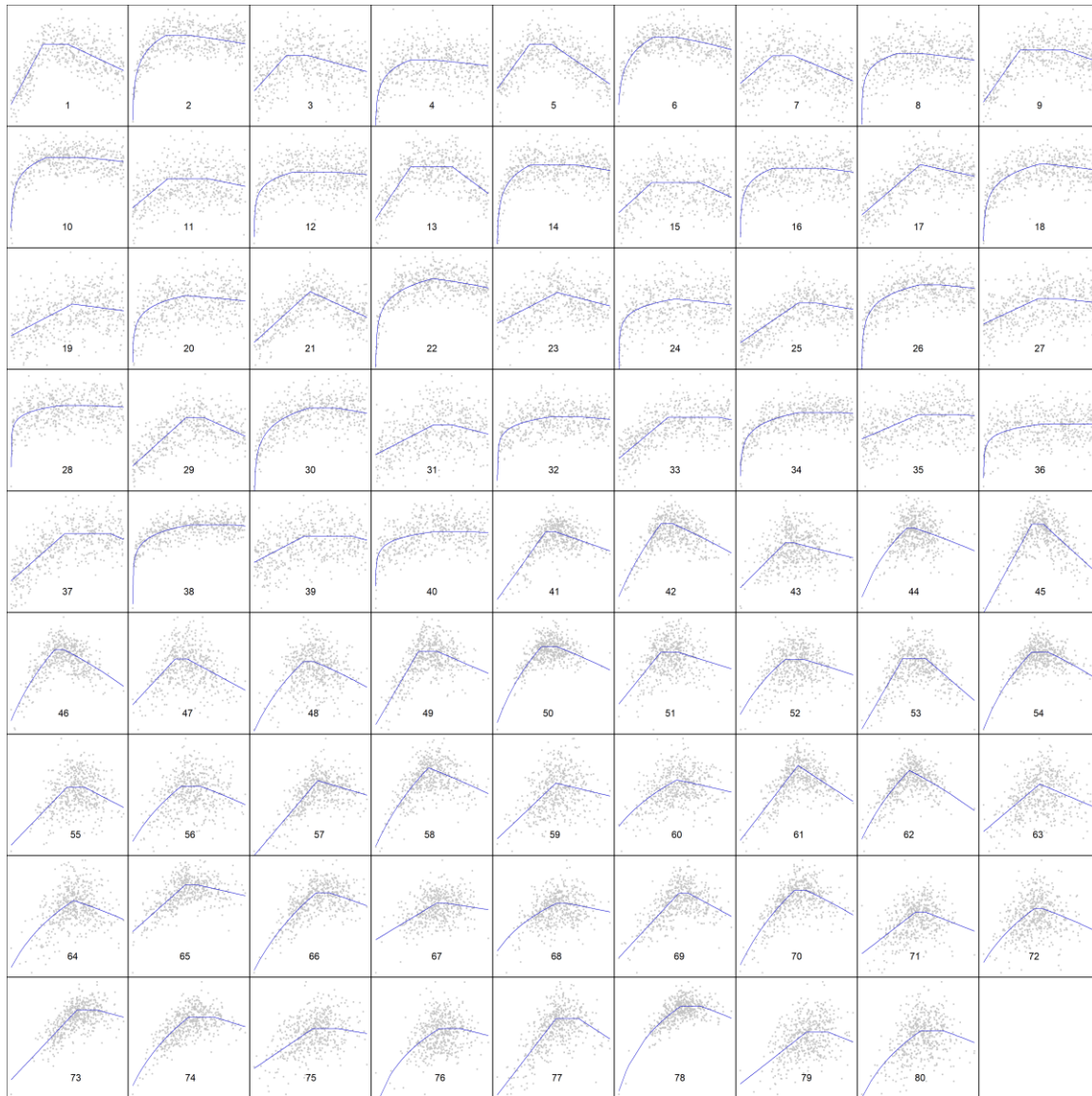R Code to reproduce simulations: https://osf.io/wdbmr, figure itself: https://osf.io/rbvw2/

Fig. 9 *A representative subset of 80 of the 2520 scenarios used to compare power across procedures to set the breakpoint in Figure 8 (left panel).* The solid lines represent the underlying true functions, the gray dots the single random draw with the specified distribution of x values and of noise. R Code to reproduce figure https://osf.io/m7avc/

Demonstrations

Figure 10 applies the two-line test to two examples in the published literature that appear

to arrive at false-positive U-shape conclusions because they relied on quadratic regressions.

Panels A&B revisit the analyses by Sterling, Jost, and Pennycook (2016) who wrote (in their

26

secondary analyses section), that "those who were moderate in terms of their support for the free market appeared to be more susceptible to bullshit than extremists in either direction." (p.356). They arrive at this inverted-U conclusion because the quadratic term in the regression is significant (p=.026; Figure 10A).

I successfully reproduced their results analyzing their posted data. Figure 10B, however, shows that the second line, while negative, is far from significant (p=.41). Keep in mind that if x and y were uncorrelated for high values of x, that is, if the true second slope were flat, 50% of estimates will be negative (and 45% of them at least as steeply as observed; that's the meaning of the p=.45). The data are inconclusive: consistent with a U-shaped relationship, consistent with ideology and bullshit receptivity being uncorrelated among higher values of the former, and consistent with a monotonic effect. Again, the U-shape prediction was secondary to the authors. The paper's core prediction is consistent with the first line in Panel B: "free market ideology was significantly but modestly associated with bullshit receptivity" (abstract).
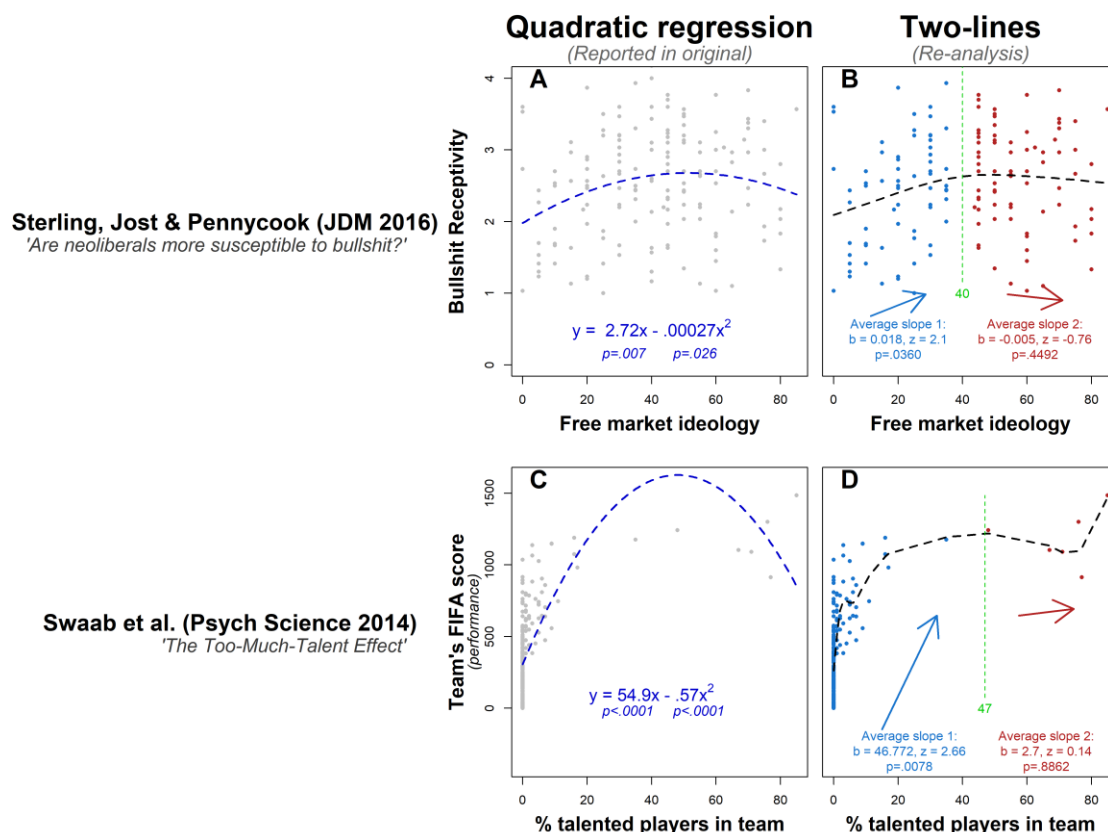
Fig. 10 *Quadratic vs Two-Lines applied to data from published papers*
Notes: In **A** & **B** each dot depicts a participant in a survey, y is how profound participants rated a series of "vague and meaningless statements," x their endorsement of "neoliberal" principles. In **C** & **D** each dot is a country, y is its FIFA rating, x the share of players in the country's team that play for a top professional team (e.g., *Arsenal*). Thin continuous lines in **B** and **D** are fitted values from cubic splines.
R Code to reproduce figure: https://osf.io/3jbzk/

Continuing with Panels C & D: Swaab, Schaerer, Anicich, Ronay, and Galinsky (2014), in their Study 2, examined the relationship between the number of elite players in a country's soccer team and its international FIFA rating. Their results, they write, "revealed a significant quadratic effect of top talent: Top talent benefited performance only up to a point, after which the marginal benefit of talent decreased *and turned negative*" (p.1584; emphasis added). I successfully replicate those results with independently obtained data (see Panel C), but in Panel D the second line is also positive (albeit far from significant). These data do not support the conclusion that there is such thing as 'too-much-talent' in international soccer.

Limitations.

Limitation 1. Asymptotic properties not examined. In this paper I have proposed an algorithm and evaluated its performance via simulation in small samples, without deriving its theoretical asymptotic properties. Moreover, the breakpoint is set with an algorithm without known theoretical properties.

Limitation 2. X, N, and W shapes. The two-lines test is expected to perform well as long as the true relationship of interest has at most two regions where the impact of x on y has opposite signs, that is, if the relationship of interest is: (i) flat overall (no effect), (ii) (weakly) monotonic, or (iii) U-shaped. It will not perform well, at least in terms of interpretability, if the true relationship has more than two regions with different signs, for instance, if it is N-shaped, X-shaped or W-shaped, rather than U-shaped. These relationships, it is worth noting, invalidate the interpretability of quadratic regressions as well. The non-parametric smooth line that accompanies the proposed app to run the test may be used as a partial solution to this limitation, alerting users if the data look N or X or W shaped.

Limitation 3. Imprecise size. The precise "size", or expected "false-positive rate", of the two-lines U-shaped test is not known, cannot be guaranteed to be 5%, for any specific dataset, for two reasons (see Supplement 8 for a detailed discussion). The first reason is that the null hypothesis of the absence of a U-shape is what is known as a "composite null." The second reason is that the Robin Hood algorithm slightly overfits. Nevertheless, the false-positive rate of the two-lines test is expected to be generally *lower* than the nominal rate, and almost never higher than 6% for a nominal $\alpha = 5\%$.

Conclusions

The use of quadratic regressions to test U-shaped relationships is as invalid as it is common. To interpret the results of a quadratic regression we need to know the true functional form is indeed quadratic, something that's virtually impossible in social science. The two-lines test of U-shaped relationships is arguably the most straightforward test of the hypothesis of interest: that the average effect of x on y is of opposite sign for high vs low values of x. It makes no assumptions about the functional form of f(x). The Robin Hood procedure to set the breakpoint for the two lines achieves notably higher power than any alternative considered. Table 1 summarizes the contents of the supplementary materials.

Table 1.
Index of supplementary materials (available from https://osf.io/6c5qb/)

Section	Pages
Supplement 1. Identifying factors that increase false-positive rate for Robin Hood	2-5
Supplement 2. Histograms with difference in power for each approach in Figure 9 in the paper.	6
Supplement 3. Two-line test with discrete x values	7-8
Supplement 4. Equivalence of Lind & Melhum (2010) with formulas in psychology textbooks, when applied to quadratic regressions.	9-11
Supplement 5. Estimating an interrupted regression does not require nor involve assuming y=f(x) is two straight lines with a discontinuity	12
Supplement 6. Two-lines vs monotonicity & nothing wrong with quadratic terms as covariates	13
Supplement 7. Accompanying quadratic regression with robustness tests is insufficient to properly use it for U-shape testing	14-15
Supplement 8. Why the false-positive rate of the two-lines test is not exactly known for any given data generating process.	16-17

References

Aiken, L. S., & West, S. G. (1991). Multiple Regression: Testing and Interpreting Interactions.

Berman, S. L., Down, J., & Hill, C. W. (2002). Tacit knowledge as a source of competitive advantage in the National Basketball Association. *Academy of Management Journal, 45*(1), 13-31.

Bowman, A., Jones, M., & Gijbels, I. (1998). Testing monotonicity of regression. *Journal of computational and Graphical Statistics, 7*(4), 489-500.

Choi, K., & Kirkorian, H. L. (2016). Touch or Watch to Learn? Toddlers' Object Retrieval Using Contingent and Noncontingent Video. *Psychological Science*. doi:10.1177/0956797616636110

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, New Jersey 07430: Lawrence Erlbaum Associates, Inc. Publishers.

Gelman, A., & Park, D. K. (2008). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician, 62*(4), 1-8.

Grant, A. M., & Schwartz, B. (2011). Too Much of a Good Thing: The Challenge and Opportunity of the Inverted U. *Perspectives on Psychological Science, 6*(1), 61-76.

Haans, R. F., Pieters, C., & He, Z. L. (2016). Thinking about U: Theorizing and testing U-and inverted U-shaped relationships in strategy research. *Strategic Management Journal, 37*(7), 1177-1195.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica, 68*(3), 575-603.

Iribarren, C., Sharp, D., Burchfiel, C. M., Sun, P., & Dwyer, J. H. (1996). Association of serum total cholesterol with coronary disease and all-cause mortality: multivariate correction for bias due to measurement error. *American journal of epidemiology, 143*(5), 463-471.

Jaspers, E., & Pieters, R. (2016). Materialism across the lifespan: An age-period-cohort analysis. *Journal of Personality and Social Psychology, 111*(3), 451-473. doi:10.1037/pspp0000092

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

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and Change in Risk-Taking Propensity Across the Adult Life Span. *Journal of Personality and Social Psychology, 111*(3), 430-450. doi:10.1037/pspp0000090

Koopmann, J., Lanaj, K., Wang, M., Zhou, L., & Shi, J. (2016). Nonlinear effects of team tenure on team psychological safety climate and climate strength: Implications for average team member performance. *Journal of Applied Psychology, 101*(7), 940-957. doi:10.1037/apl0000097

Lind, J. T., & Mehlum, H. (2010). With or Without U? The Appropriate Test for a U-Shaped Relationship. *Oxford Bulletin of Economics and Statistics, 72*(1), 109-118.

Loschelder, D. D., Friese, M., Schaerer, M., & Galinsky, A. D. (2016). The Too-Much-Precision Effect When and Why Precise Anchors Backfire With Experts. *Psychological Science*. doi:10.1177/0956797616666074

Marsh, L. C., & Cormier, D. R. (2001). *Spline regression models* (Vol. 137): Sage.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological methods, 2*(1), 3-19.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological bulletin, 114*(2), 376.

Miller, J. W., Stromeyer, W. R., & Schwieterman, M. A. (2013). Extensions of the Johnson-Neyman technique to linear models with curvilinear effects: Derivations and analytical tools. *Multivariate behavioral research, 48*(2), 267-300.

Molinari, N., Daures, J.-P., & Durand, J.-F. (2001). Regression splines for threshold selection in survival data analysis. *Statistics in medicine, 20*(2), 237-247.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in medicine, 22*(19), 3055-3071.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General, 145*(10), 1269-1279.

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.

Qian, G., Khoury, T. A., Peng, M. W., & Qian, Z. (2010). The performance implications of intra- and inter-regional geographic diversification. *Strategic Management Journal, 31*(9), 1018-1030.

Seidman, G. (2012). Positive and negative: Partner derogation and enhancement differentially related to relationship satisfaction. *Personal Relationships, 19*(1), 51-71.

Simonsohn, U., & Nelson, L. D. (2014). DataColada [27] - Thirty-somethings are shrinking and other u-shaped challenges. Retrieved from http://www.datacolada.com/27

Simonton, D. K. (1976). Biographical determinants of achieved eminence: a multivariate approach to the Cox data. *Journal of Personality and Social Psychology, 33*(2), 218.

Spiller, S. A., Fitzsimons, G. J., Lynch Jr, J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research, 50*(2), 277-288.

Stasinopoulos, D., & Rigby, R. (1992). Detecting break points in generalised linear models. *Computational Statistics & Data Analysis, 13*(4), 461-471.

Sterling, J., Jost, J. T., & Pennycook, G. (2016). Are neoliberals more susceptible to bullshit? *Judgment and Decision Making, 11*(4), 352-360.

Swaab, R. I., Schaerer, M., Anicich, E. M., Ronay, R., & Galinsky, A. D. (2014). The too-much-talent effect team interdependence determines when more talent is too much or not enough. *Psychological Science, 25*(8), 1581-1591.

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science, 22*(2), 253-260.

von Bastian, C. C., Souza, A. S., & Gade, M. (2016). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of Experimental Psychology: General, 145*(2), 246-258.

Wilson, K. S., DeRue, D. S., Matta, F. K., Howe, M., & Conlon, D. E. (2016). Personality similarity in negotiations: Testing the dyadic effects of similarity in interpersonal traits and the use of emotional displays on negotiation outcomes. *Journal of Applied Psychology, 101*(10), 1405-1421. doi:10.1037/apl0000132

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