Non-Robustness in Log-Like Specifications

Jack Fitzgerald, Joop Adema, Lenka Fiala, Essi Kujansuu, and David Valenta

March 13, 2026

Logarithms and Percentage Effects

Researchers are often interested in estimating percentage effects and (semi-)elasticities

- **Standard statistical training:** Before linear regression, transform variables of interest with the natural logarithm function $\ln(Z)$

Challenge: $\ln(Z)$ is undefined for non-positive Z

- Logarithmic specifications are inestimable for variables with non-positive values unless those values are dropped
- Undesirable due to power loss and sample selection concerns

$\ln(Z)$ being undefined at $Z = 0$ is fundamentally related to the credibility of logarithmic specifications for estimating percentage effects

- The percentage change from zero to any other real number is undefined, so of course $\ln(0)$ is undefined as well

'Log-Like' Transformations

There are some 'log-like' transformations that asymptotically approach $\ln(Z)$ as $Z \to \infty$, but are defined at zero Formal Definition

- Researchers using these transformations for percentage effect estimation effectively argue that they can return valid percentage effect estimates even with zeros in the data (see Bellemare & Wichman 2020)

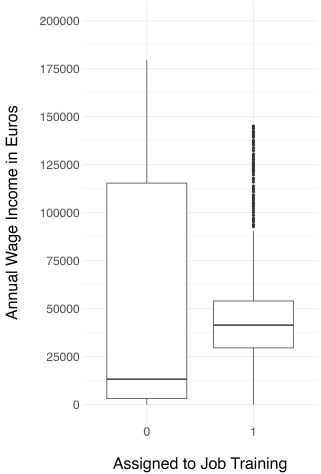
Common examples include the $\ln(Z + c)$ and inverse hyperbolic sine (IHS) transformation

$$\sinh^{-1}(Z) = \ln\left(\sqrt{1 + Z^2} + Z\right)$$

Usage of these transformations (in particular the IHS) has been popularized by several influential methodological recommendations with hundreds of citations (Burbidge, Magee, & Robb 1988 [799 Web of Science citations; 1882 Google Scholar citations]; Bellemare & Wichman 2020 [665 Web of Science citations; 1607 Google Scholar citations])

- Since 2021, 3-4% of all publications in *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics* have discussed IHS specifications (Goldsmith-Pinkham 2024)

Motivating Example



Suppose (hypothetically) that the EU Joint Research Centre runs an EU-wide RCT to evaluate the effects of an online job training program on income

- We simulate experimental data on 1000 participants (distribution to the left)
- Some individuals are unemployed and thus receive zero wage income

Annual wage income is principally measured in euros

- However, would also be reasonable to measure in 1000s of euros or euro cents
- Euros also aren't the only currency in the EU; could measure in Hungarian forint, Swedish krona, etc.

Scale Variance

Unlike differences in logs, differences in log-like transformations are not scale-invariant

- This means regression coefficients in log-like specifications are scale-variant

	(1) Euros (Raw)	(2) Euros (1000s)	(3) Euros (Cents)	(4) Czech Koruna	(5) Danish Krone	(6) Hungarian Forint	(7) Polish Złoty	(8) Romanian Leu	(9) Swedish Krona
Panel A: $\ln(aY)$									
Treatment	1.155***	1.155***	1.155***	1.155***	1.155***	1.155***	1.155***	1.155***	1.155***
	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)
N	922	922	922	922	922	922	922	922	922
Panel B: IHS(aY)									
Treatment	-0.610***	0.432***	-1.329***	-1.108***	-0.924***	-1.537***	-0.834***	-0.864***	-0.978***
	(0.208)	(0.115)	(0.276)	(0.255)	(0.237)	(0.297)	(0.228)	(0.231)	(0.242)
N	1000	1000	1000	1000	1000	1000	1000	1000	1000
Scaling Parameter	$a = 1$	$a = 0.001$	$a = 100$	$a = 24.267$	$a = 7.4677$	$a = 380.6$	$a = 4.2023$	$a = 5.0965$	$a = 10.5820$

HC3 heteroskedasticity-robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Inverse Hyperbolic Sines

Chen & Roth (2024): In regressions involving log-like variables and zeros in the data, any coefficient magnitude can be obtained by re-scaling log-like-transformed variables

- **Thakral & Tô (2025):** Test statistics in log-like specifications are similarly sensitive to the scale of log-like-transformed variables

Implies that log-like specifications may be quite non-robust in practice

- This project examines the degree to which these specifications threaten published research

Empirical Contribution

We revisit replication data of 46 papers with main findings defended by log-like specifications

- Holding all other specification choices constant, we examine the effect of several re-scaling and functional form adjustments on results

Simply removing the log-like transformation changes conclusions for 36% of estimates

- Different re-scalings change conclusions for 16-30% of estimates

99.8% of estimates neglect data recommendations from methodological guidelines on log-like specifications (Bellemare & Wichman 2020)

- For 21% of estimates, at least one main log-like variable's input has no non-positive values

Stat. sig. results are $> 40\%$ more frequent in log-like specifications than in the economics literature

- For 38% of estimates, inputs to log-like transformations are scaled to a *sweet spot*; both above and below this scaling, t-statistics are weaker
- The stat. significance of sweet spot estimates is significantly more sensitive to specification choice

Theoretical Contribution

Our replication results motivate the development of new theoretical results concerning the behavior of test statistics in log-like specifications

- Test statistics in log-like specifications often exhibit 'sweet spots' (i.e., local optima) in unit scale
- For narrow bands of unit scales, test statistics in log-like specifications can briefly dip into rejection regions that change statistical significance conclusions
- Even in simulated data with no treatment effect, one can still achieve rejection rates in log-like specifications well above nominal significance levels by searching over different unit scalings
- Our simulation results show that this is an overfitting problem: the unit scalings yielding the most spuriously significant results also yield the worst out-of-sample prediction performance

We harmonize the literature's recommendations for more robust alternative specifications

- We specifically highlight normalized estimands, Poisson regression, and quantile regression as good options

Background and Theory

Log-Like Transformations Formalized

Chen & Roth (2024) deem a transformation $m(Z)$ **log-like** if it satisfies two properties:

1. $m(Z)$ asymptotically converges to $\ln(Z)$:

$$\lim_{Z \to \infty} \frac{m(Z)}{\ln(Z)} = 1$$

2. $m(Z)$ is defined at zero

Most popular transformations satisfying these conditions include the $\ln(Z+1)$ transformation and the IHS transformation

$$\sinh^{-1}(Z) = \ln\left(\sqrt{1+Z^2} + Z\right)$$

Back to 'Log-Like' Transformations

Notation

Without loss of generality, we'll consider **log-like specifications** where outcome $Y \geq 0$ and only Y is transformed:

$$m(Y) = X\beta_{\text{LL}} + \epsilon$$

One could instead parameterize this as a **linear** specification:

$$Y = X\beta_{\text{Lin}} + \mu$$

Alternatively, one could estimate an **extensive-margin** specification:

$$\mathbb{1}[Y > 0] = X\beta_{\text{EM}} + \eta$$

Suppose that original outcome Y can be linearly scaled by some $a > 0$ (before transformation)

- Let $\beta_{\text{LL}}(a)$ indicate the regression parameter β_{LL} estimated for transformed outcome $m(aY)$, and $t_{\text{LL}}(a)$ be its t-statistic

How Log-Like Coefficients Vary with Scale

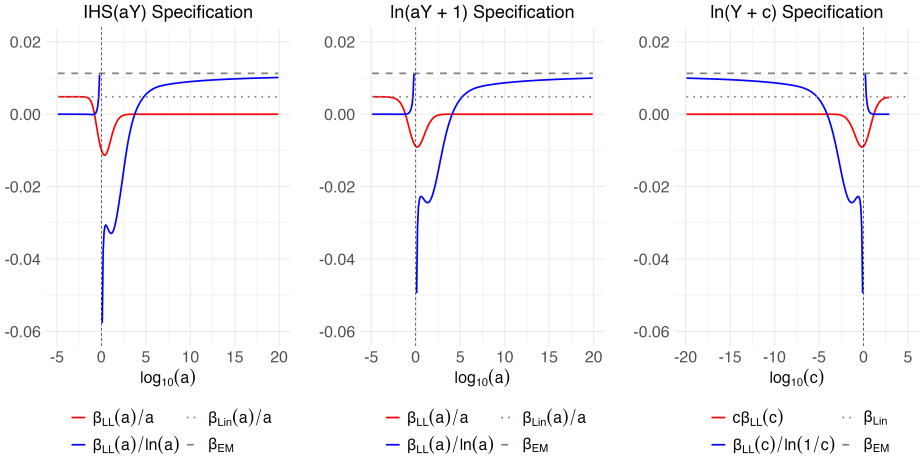
Chen & Roth (2024) show that $\lim_{a \to \text{inf}} \beta_{\text{LL}}(a) \approx \ln(a) \times \beta_{\text{EM}}$

- **Problem:** $\lim_{a \to \text{inf}} \ln(a) = \infty$, so $\beta_{\text{LL}}(a)$ can be scaled infinitely large by just scaling up Y (unless there is no extensive margin, and thus $\beta_{\text{EM}} = 0$)
- Intuitively, $\lim_{a \to 0} \beta_{\text{LL}}(a) = 0$ because as $a \to 0$, so too does all variation in $m(aY)$

Because $\beta_{\text{LL}}(a)$ is continuous for all $a > 0$, by intermediate value theorem, a researcher can set $\beta_{\text{LL}}(a)$ to **any desired size** just by scaling Y

- This should not be possible for a consistent percentage effect or (semi-)elasticity estimator, as these parameters are not scale-variant

Asymptotic Behavior of Coefficients in Log-Like Specifications



Constants in $\ln(aY + c)$ Specifications

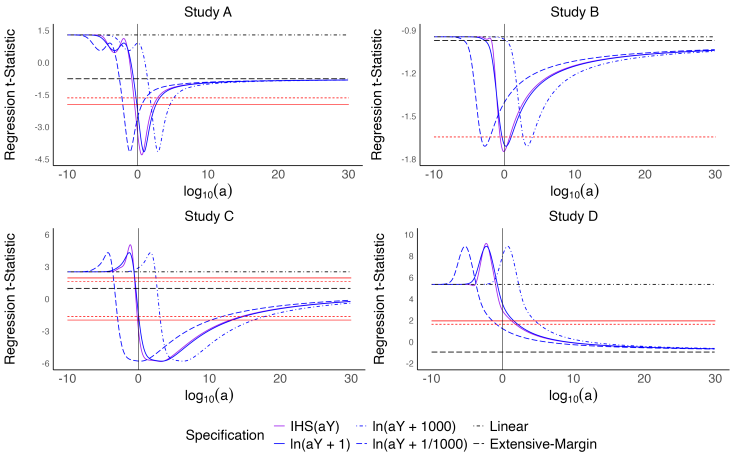
Mullahy & Norton (2024) highlight that in $\ln(aY + c)$ specifications, coefficients are sensitive to c in the same way that they are sensitive to a

- Intuitively, this is because with zeros in the data, log-like transformations require some parameterization of the distance between zero and non-zero values
- In the $\ln(aY + c)$ transformation, non-zeros are pushed further away from zeros both when a increases and when c decreases
- This is why the second and third graphs in the previous figure are mirror images over the y-axis

Chen & Roth (2024) show that you can get arbitrarily large coefficients in $\ln(aY + c)$ specifications by making a arbitrarily large

- It follows that you can also get arbitrarily large coefficients by making c arbitrarily small

Test Statistic Behavior in Log-Like Specifications



Our replications showed:

- $t_{LL}(a)$ often hits peaks and troughs outside the convex hull of t_{Lin} and t_{EM}
- These peaks + troughs are often sweet spots at which $t_{LL}(a)$ only briefly dips into rejection regions
- In $\ln(aY + c)$ specifications, the scale-variance of $t_{LL}(a)$ is shift-variant with respect to c Version for c

Why Are Sweet Spots Possible?

Regression t-statistics are ratio statistics of the form $\hat{\beta}/\text{SE}\left(\hat{\beta}\right)$; when scale can affect t-statistics,

$$\frac{\partial t(a)}{\partial a} = \frac{\text{SE}\left(\hat{\beta}(a)\right) \times \frac{\partial \hat{\beta}(a)}{\partial a} - \hat{\beta}(a) \times \frac{\partial \text{SE}\left(\hat{\beta}(a)\right)}{\partial a}}{\text{SE}\left(\hat{\beta}(a)\right)^2}$$

Sweet spots can form whenever $t(a)$ is **non-monotonic** in a

- Monotonicity of t-statistics requires a global inequality constraint over the first and second terms in the numerator for all values of a
- Proving that t-statistics can be non-monotonic with respect to scale requires only one empirical counterexample to this inequality constraint

Our replications reveal dozens of counterexamples across numerous published articles

Why Aren't All t-Statistics Scale-Variant?

When estimands are *scale-equivariant* (e.g., coefficients from standard linear OLS specifications):

$$\frac{\hat{\beta}(a)}{\text{SE}\left(\hat{\beta}(a)\right)} = \frac{a \times \hat{\beta}(1)}{a \times \text{SE}\left(\hat{\beta}(1)\right)} = \frac{\hat{\beta}(1)}{\text{SE}\left(\hat{\beta}(1)\right)}$$

When estimands are *scale-invariant* (e.g., coefficients from standard logarithmic OLS specifications):

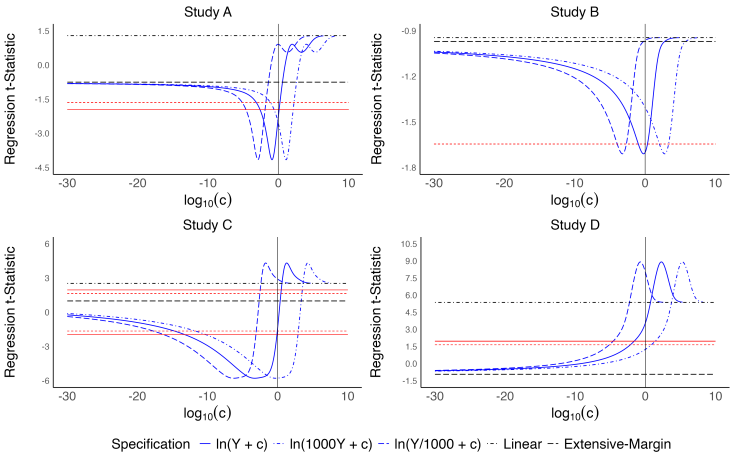
$$\frac{\hat{\beta}(a)}{\text{SE}\left(\hat{\beta}(a)\right)} = \frac{\hat{\beta}(1)}{\text{SE}\left(\hat{\beta}(1)\right)}$$

In both cases, scale effectively does not enter the t-statistic, and thus

$$\frac{\partial t(a)}{\partial a} = 0$$

This guarantees a stability of test statistics which log-like specifications do not possess

Shift-Variance of Test Statistics



Our replications also showed:

- In $\ln(aY + c)$ specifications, sweet spots emerge not just in a, but also in c
- $t_{\text{LL}}(c)$ exhibits sweet spots and escapes the convex hull in a similar way as we observe for a
- In fact, the t-curves in a and c are essentially mirror images of one another over the y-axis

Version for a

Why Should We Care?

With scale-variant t-statistics, error rates can be completely uncontrolled due to multiple testing

- There are literally an infinite number of tests to run which are equally theoretically valid, most of which provide different results, and some of which yield different conclusions

Consider the following scenarios:

- A single researcher estimates the same relationship, mining over scalings to optimize statistical significance (p-hacking)
- Many researchers estimate the same relationship, each using a different scaling, and only those with statistically significant results submit manuscripts (supply-side publication bias)
- Many researchers estimate the same relationship, each using a different scaling, and all submit manuscripts, but journals only publish the manuscripts with statistically significant results (demand-side publication bias)

Each of these scenarios highlight a pathway by which the scale-variance of t-statistics can allow spuriously significant results to spill into the literature

The Rise and Fall of Log-Like Specifications

Log-like specifications were popularized in economics by Bellemare & Wichman (2020), which recommended IHS specifications for (semi-)elasticity estimation when there are zeros in the data

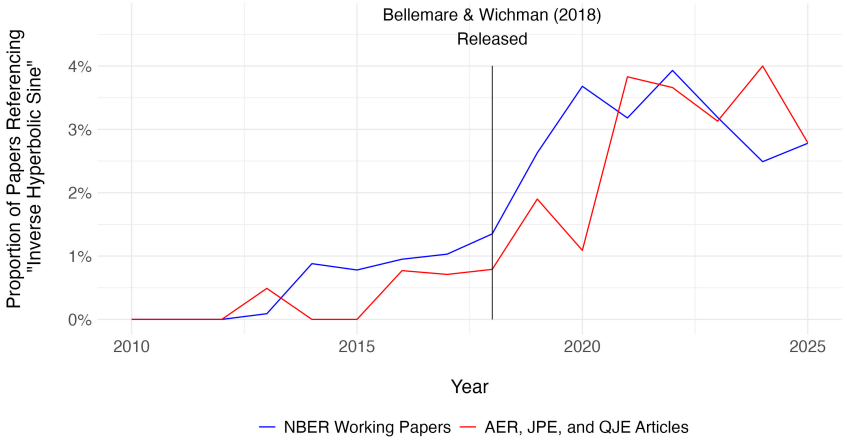
- The paper has 1607 Google Scholar citations as of 12 March 2026
- Since 2021, 3-4% of all publications in *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics* have discussed IHS specifications (Goldsmith-Pinkham 2024)
- Norton (2022) subsequently published a Stata tutorial on computing marginal effects in IHS specifications

Shortly after Bellemare & Wichman (2020) was published, numerous critiques of log-like specifications were published in rapid succession (Aihounton & Henningsen 2021; Cohn, Liu, & Wardlaw 2022; Mullahy & Norton 2024; Chen & Roth 2024; Thakral & Tô 2025)

- Each theoretically documents the scale-sensitivity of log-like specifications and offers recommendations (which sometimes conflict with one another)

These critiques do not appear to have meaningfully slowed the growth of log-like specifications

Log-Like Specifications in Economics



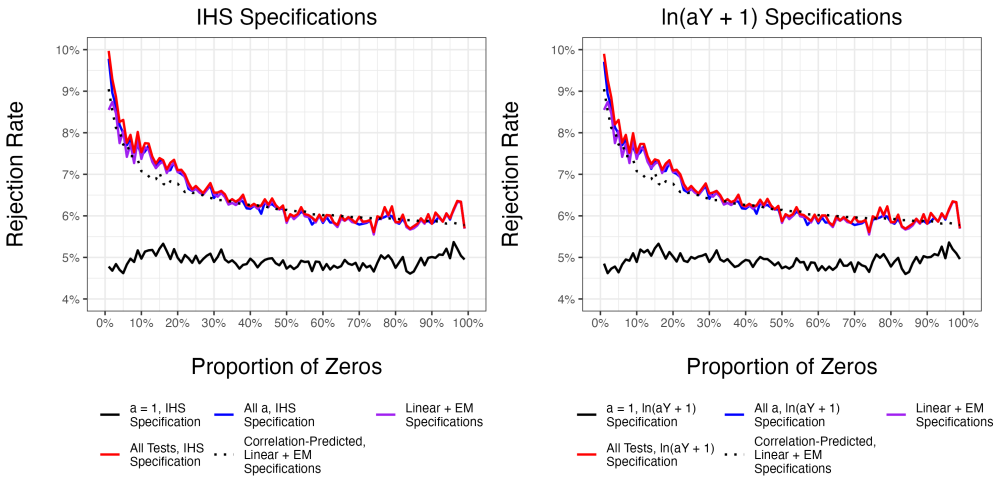
Source: paulgp.com/econlit-pipeline (see Goldsmith-Pinkham 2024)

Simulation Evidence

Setup

- Draw samples of Y values from $N(25, 5)$ and randomly assign half of the sample to a binary placebo treatment X
- Set some proportion $p_0 \in \{0.01, 0.02, \cdots 0.99\}$ of the Y values to zero
- For each p_0, we conduct 10,000 draws of 22,900 observations each
- Estimate IHS, linear, extensive-margin, and $\ln(aY + 1)$ specifications for $a \in \{10^{-7}, 10^{-6.9}, \cdots 10^{10}\}$

Rejection Rates

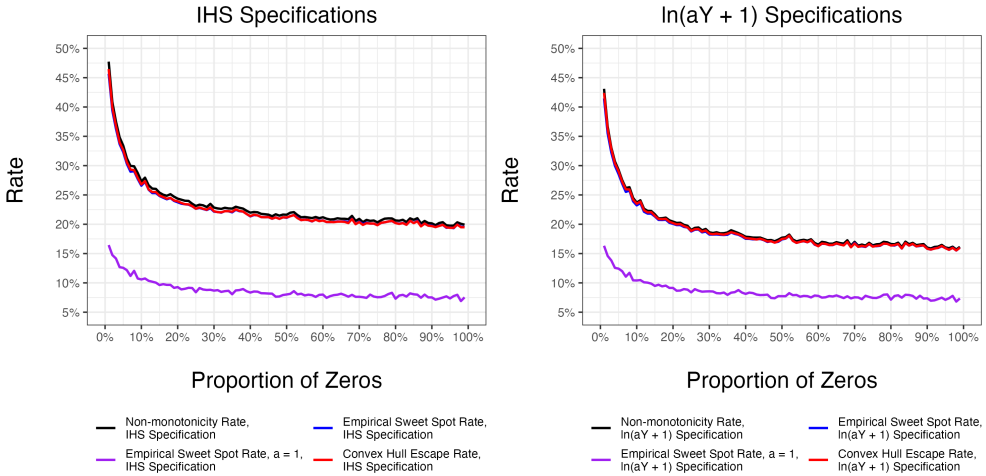


Rejection Rates Explained

Definition of rejection rate: The proportion of simulation draws in which the null hypothesis of no treatment effect is rejected at the 5% significance level, using a two-sided t-test with heteroskedasticity-robust standard errors

- When we fix $a = 1$, rejection rates remain controlled near the 5% nominal rate
- When a researcher can pick any scaling a, rejection rates increase to 6.43% on average
- Similar to (but slightly higher than) the rejection rate you get if the researcher is allowed to report the more significant between the linear and extensive-margin specifications
- Slightly higher, because $t_{\text{LL}}(a)$ can escape the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}
- The lower the share of zeros in the data (p_0), the higher the rejection rate

Non-monotonicity Rates



Non-monotonicity Rates Explained

- $t_{LL}(a)$ is non-monotonic in a in 21.12% of all specifications
- We observe 'empirical sweet spots' whenever $t_{LL}(a)$ is larger than both $t_{LL}\left(10^{-3}a\right)$ and $t_{LL}\left(10^{3}a\right)$; this occurs for $a = 1$ in 8.61% of specifications and for some a in 20.7%
- For some a, $t_{LL}(a)$ escapes the convex hull of $t_{Lin}(a)$ and t_{EM} in 20.74% of all specifications
- The lower the share of zeros in the data, the higher these non-monotonicity rates become

Explaining Simulation Results: Overfitting

Why are rejection and non-monotonicity rates so much higher for lower p_0? **Overfitting**

- Spurious significance arises because for some a, $\mathbb{E}[m(aY) \mid X]$ fits our simulated noise better
- The more dense the intensive margin, the easier it is to find noise to fit

How do we know?

- For each of 500 draws, we split the data into training and testing data
- For each draw Δ, we estimate IHS and $\ln(aY+1)$ specifications s within both the training and testing datasets and store regression-level results
- We then regress $100R^2_{\Delta,s,a} = \chi + \beta \text{WithinSampleMetric}_{\Delta,s,a} + \pi_{\Delta,s} + \upsilon_{\Delta,s,a}$
- $100R^2_{\Delta,s,a}$ is R^2 either within-sample or out-of-sample measured in p.p. units
- Four within-sample metrics Details
- Draw-by-specification fixed effects $\pi_{\Delta,s}$

Skip Ahead - Connection to Model Selection Criteria

More Spuriously Significant Results Have Higher *Within-Sample* R^2...

	(1)	(2)	(3)	(4)
Dependent Variable: Within-Sample R^2				
Within-Sample $\lvert t_{LL}(a)\rvert$	0.015898			
	(0.00007)			
Within-Sample Statistical Significance		0.013202		
		(0.000185)		
Within-Sample Empirical Sweet Spot			0.000678	
			(0.000011)	
Within-Sample Convex Hull Escape				0.00018
				(0.000014)
N	16,929,000	16,929,000	16,483,526	16,929,000
# Clusters	99,000	99,000	99,000	99,000

... But Lower *Out-of-Sample* R^2

	(1)	(2)	(3)	(4)
Dependent Variable: Out-of-Sample R^2				
Within-Sample $\lvert t_{LL}(a) \rvert$	-0.016151			
	(0.000251)			
Within-Sample Statistical Significance		-0.012535		
		(0.000391)		
Within-Sample Empirical Sweet Spot			-0.000667	
			(0.000038)	
Within-Sample Convex Hull Escape				-0.000158
				(0.000048)
N	16,929,000	16,929,000	16,483,526	16,929,000
# Clusters	99,000	99,000	99,000	99,000

Data and Methods

Data

Papers don't typically advertise that they use log-like transformations in searchable fields

- To find papers using these transformations, our starting sample is the 423 published articles recorded by Web of Science as citing Bellemare & Wichman (2020) as of 17 August 2024

We revisit publicly-available and suitable replication data for 46 papers with at least one abstract-level claim defended by a log-like specification What is a claim?

- For 97% of estimates, the outcome/exposure of interest is IHS-transformed; the remainder use $\ln(Z + c)$ specifications

This yields 596 estimates defending 137 claims

- For each estimate, we store original specification information and attempt to reproduce the original results (we perfectly match published results 80% of the time)
- For each specification (reproducibility and various robustness), we extract the relevant regression coefficient, standard error, and p-value

Functional Form Adjustments

Our first adjustment is *linear* retransformation:

$$m(Y_i) = v + \sum_{\ell=1}^{k_L} \beta_\ell m(X_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i \quad \Rightarrow \quad Y_i = v + \sum_{\ell=1}^{k_L} \beta_\ell X_{i,\ell} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i$$

We supplement this with a nonlinear *cube root* retransformation (domain-preserving and concave across entire domain) Functional Forms

$$m(Y_i) = v + \sum_{\ell=1}^{k_L} \beta_\ell m(X_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i \quad \Rightarrow \quad \sqrt[3]{Y_i} = v + \sum_{\ell=1}^{k_L} \beta_\ell \sqrt[3]{X_{i,\ell}} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i$$

Finally, for the 55.5% of estimates where they're estimable, we run Poisson specifications:

$$m(Y_i) = v + \sum_{\ell=1}^{k_L} \beta_\ell m(X_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i \quad \Rightarrow \quad Y_i = \exp\left(v + \sum_{\ell=1}^{k_L} \beta_\ell Z_{i,\ell} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i \right)$$

Scaling Adjustments

We conduct three re-scalings:

$$m(Y_i) = v + \sum_{\ell=1}^{k_L} \beta_\ell m(X_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i \quad \Rightarrow \quad m(aY_i) = v + \sum_{\ell=1}^{k_L} \beta_\ell m(aX_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i$$

1. mul1000: $a = 1000$
2. div1000: $a = 1/1000$
3. min10: $a = 10/\min_{Z_i \neq 0}(|Z_i|)$

Though mul1000 and div1000 are arbitrary transformations, min10 has roots in recent methodological recommendations (Bellemare & Wichman 2020)

Estimating Specification Effects

Because we vary specifications *ceteris paribus* within the same estimate, we can construct an estimate-specification panel dataset

- A row of our data is the results of specification s for estimate i

We code two measures of robustness, both at the 5% and 10% significance levels: Details

1. $\text{Agree}_{i,s}$: Indicates whether specification s yields the same tripartite statistical significance conclusion as the reproduction specification for estimate i
2. $\text{Sig}_{i,s}$: Indicates whether specification s is statistically significantly different from zero

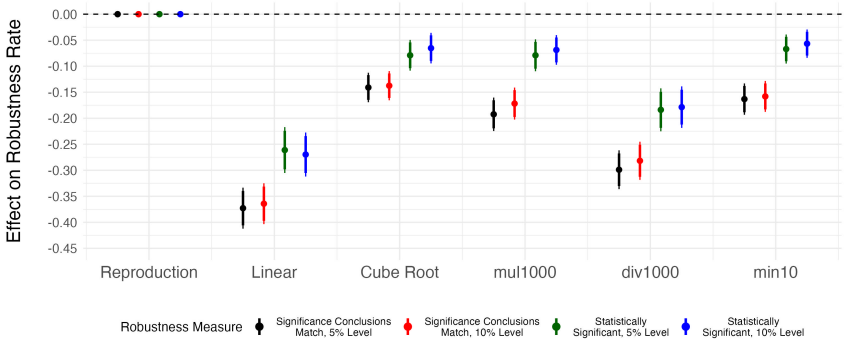
We estimate effects of different specifications on robustness with fixed effects models:

$$\text{Sig}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}$$

Because we hold all other parts of the estimation process constant, γ_s identifies the effect of specification choice s on robustness

Replication Results

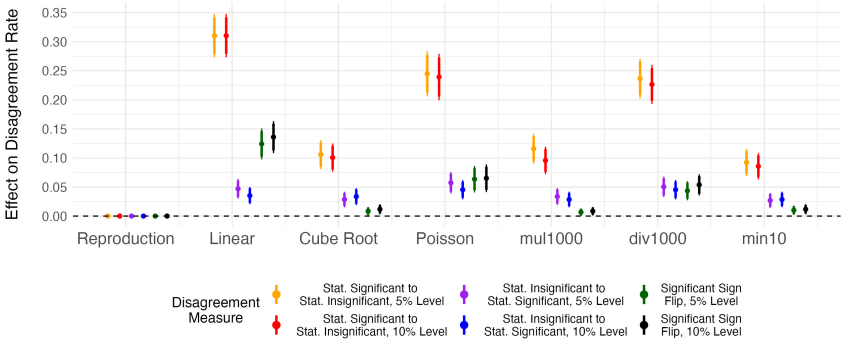
Main Results



Our robustness checks change significance conclusions for 14-37% of estimates

- Compared to reproduction estimates, the rate of statistically significant results in our robustness specifications decreases by 6-28 p.p.

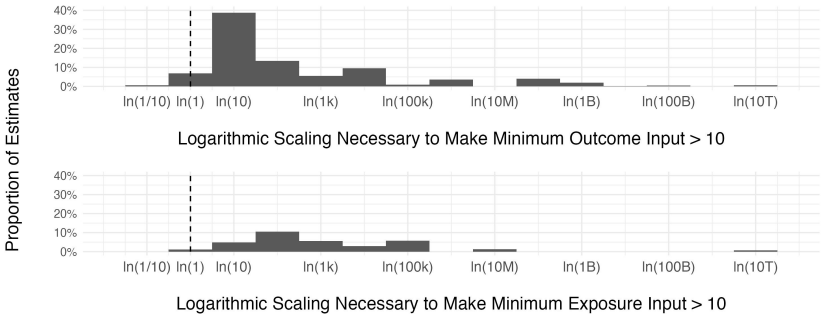
Decomposing Agreement Effects



Specification effects on conclusion agreement are predominantly (though certainly not entirely) driven by statistically significant results losing statistical significance after robustness checks

- For 12% of estimates, both the reproduction and linear specifications are statistically significant, but take opposite signs

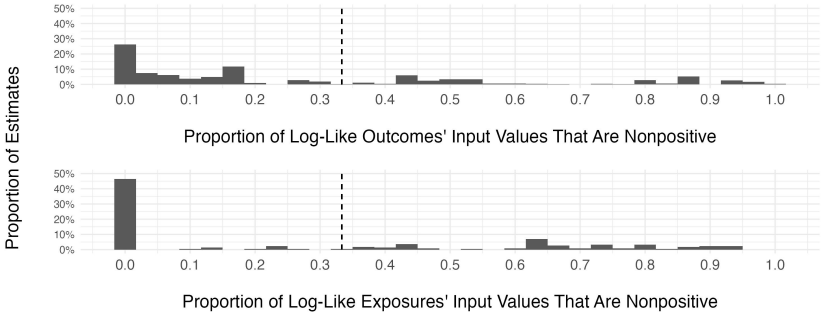
Adherence to Methodological Recommendations on Unit Scale



Because IHS(Z) and ln(Z) are only close for large values of Z, Bellemare & Wichman (2020) recommend that IHS only be used for variables with minimum non-zero values ≥ 10

- For 99.8% of estimates, at least one variable transformed with log-like functions neglects Bellemare & Wichman's (2020) recommendation that min $\{|Z| \mid Z \neq 0\} \geq 10$

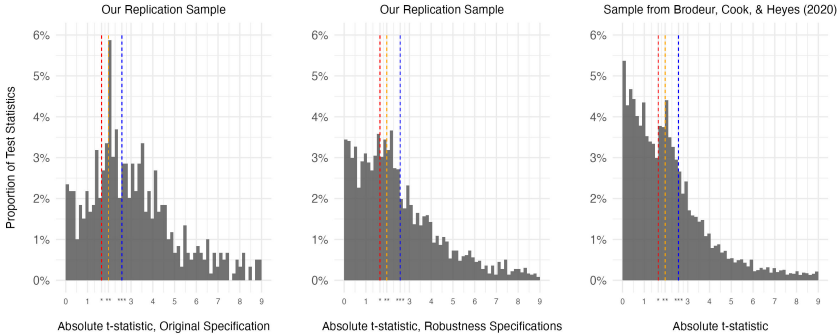
Proportions of Non-positives



Bellemare & Wichman (2020) recommend that $\leq 1/3$ of Z's values be zeros before IHS transformation; for 32-38% of estimates, ≥ 1 variable transformed with log-like functions neglects this

- 13% (41%) of outcomes (exposures) of interest transformed with log-like functions have *zero* non-positive values, leaving no justification for log-like transforms

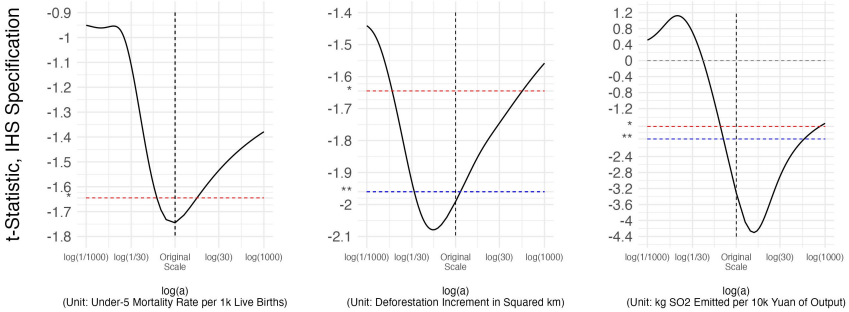
Publication Bias



Huge mass of reproduction test statistics right beyond the 5% statistical significance threshold, which is unobserved in both our robustness specifications and the causal economics literature

- Compared to the causal economics literature, log-like specifications yield statistically significant results 49% (40%) more frequently at the 5% (10%) level

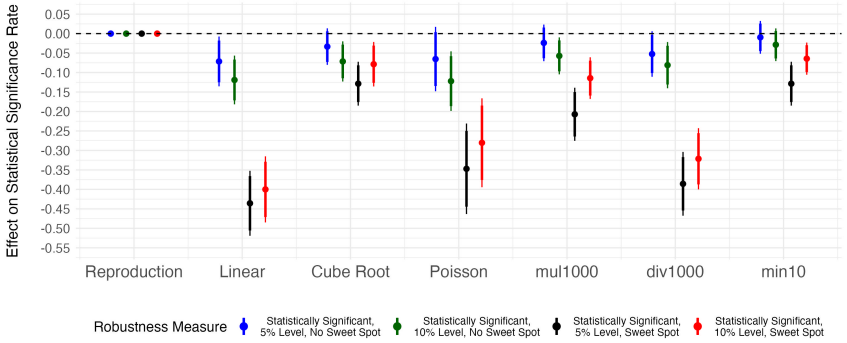
The Sweet Spot



38% of estimates are in the sweet spot: the reproduction *t*-statistic is larger than *both* the mul1000 and div1000 *t*-statistics

- For 8% of estimates, *both* the mul1000 and div1000 specifications yield different conclusions than the reproduction specification

Heightened Sensitivity of Sweet Spot Estimates



The (5%) stat. significance of sweet spot estimates is 14-36 p.p. more sensitive to specification choice than estimates outside the sweet spot

- Specification effects on stat. significance are only robustly stat. sig. diff. from zero for estimates in the sweet spot

Recommendations

What Doesn't Work: Power Specifications

Thakral & Tô (2025) recommend power transforms $g(Z) = Z^{\omega}$ $(\omega > 0)$, as power specifications yield scale-equivariant estimands even with zeros in the data

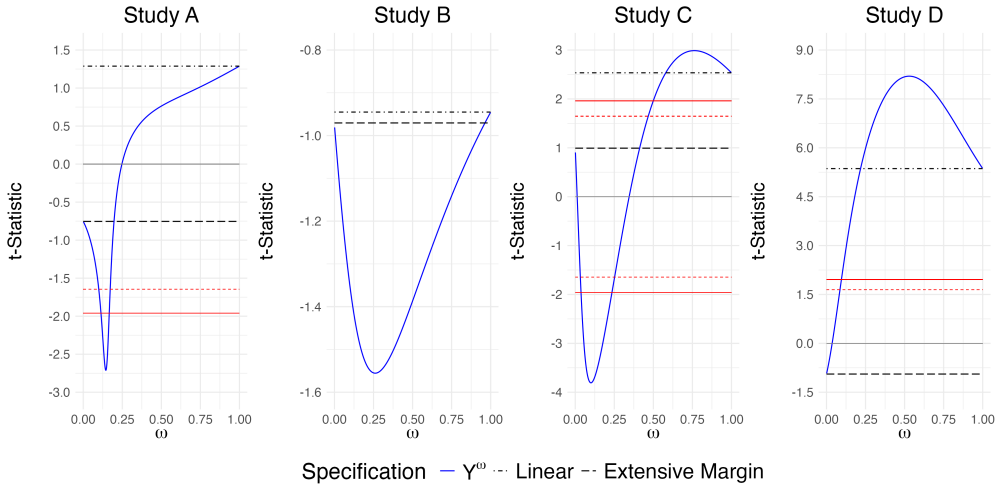
- However, Mullahy & Norton (2024) recommend against power specifications because their estimands are sensitive to the choice of ω, which is theoretically arbitrary
- We concur with this latter perspective

For concave power transforms, power specifications exhibit the same convergence properties as log-like specifications

- As $\omega \to 1$, power specifications trivially converge to linear specifications
- As $\omega \to 0$, all differences between non-zero values get flattened, leaving extensive-margin variation as the only variation

As in log-like specifications, test statistics in power specifications also exhibit sweet spots in ω

Test Statistic Behavior in Power Specifications



What Doesn't Work: Model Selection Criteria

Aihounton & Henningsen (2021) recommend estimating log-like specifications for different scaling parameters a and finding the a which optimizes some model selection criterion

- Aihounton & Henningsen (2021) recommend 14 different potential model selection criteria, but generally prefer R^2

There are two issues with this recommendation

1. It is difficult to restrict researchers choosing from among model selection criteria/scaling granularities that optimize statistical significance
2. Finding the scaling which happens to make log-like transformations fit the data best is precisely the kind of specification searching that drives spurious significance, as our simulation evidence shows Back to Explaining Simulation Results: Overfitting

What Doesn't Work: Extensive-Margin Calibration

Chen & Roth (2024) propose using 'calibrated extensive-margin' transformations of the form

$$m(Y, \psi) = \begin{cases} \ln(Y) \text{ if } Y > 0 \\ -\psi \text{ if } Y = 0 \end{cases},$$

where the researcher specifies ψ themselves

- In this strategy, the researcher themselves 'calibrates' (i.e., chooses) what the value of the extensive margin effect is

This does not resolve the fundamental credibility issues with log-like specifications

- Resulting specifications still yield t-statistics that are scale-variant and exhibit Sweet Spots in a
- These t-statistics also vary with ψ and exhibit Sweet Spots in ψ
- By sending $\psi \to \infty$, the researcher can still put infinite weight on the extensive-margin effect
- If $-\psi \geq \ln\left(\min_{Y>0} Y\right)$, then $m(Y, \psi)$ assigns (weakly) higher values to zeros than to some positives

What Doesn't Work: Two-Part Models

Bellemare & Wichman (2020) and Mullahy & Norton (2024) recommend two-part models as a (potential) alternative, which estimate:

$$\tau_a = P(Y(1) > 0 \mid D = 1) - P(Y(0) > 0 \mid D = 0)$$
$$\tau_b = \mathbb{E}[Y \mid Y > 0, D = 1] - \mathbb{E}[Y \mid Y > 0, D = 0]$$

Notice that τ_b is a parameter *conditioned* on $Y > 0$

- **Problem:** D can affect whether $Y > 0$

Conditioning on $Y > 0$ creates a 'bad control problem', which can induce collider biases (Cinelli, Forney, & Pearl 2024)

- Chen & Roth (2024) show that τ_b can be decomposed as a causal intensive margin effect and (possibly multiple) selection effects
- These existence of these selection effects is what motivates the use of log-like specifications in the first place

What Doesn't Work: Lee Bounds

Chen & Roth (2024) recommend using Lee (2009) bounds to estimate partially-identified intensive-margin relationships:

$$\hat{\beta}^{+}_{\text{Lee}} = \mathbb{E}\left[\ln(Y) \mid X = 1,\ Y > 0,\ Y \geq Y_q\right] - \mathbb{E}\left[\ln(Y) \mid X = 0,\ Y > 0\right]$$
$$\hat{\beta}^{-}_{\text{Lee}} = \mathbb{E}\left[\ln(Y) \mid X = 1,\ Y > 0,\ Y \leq Y_{1-q}\right] - \mathbb{E}\left[\ln(Y) \mid X = 0,\ Y > 0\right]$$

Problem: Partial identification requires a monotonicity assumption that all observations with $Y > 0$ and $X = 0$ would also have $Y > 0$ if it were instead the case that $X = 1$

- A purely unidirectional effect of exposure on $\Pr(Y > 0)$ is unlikely in practice

Researchers can also tighten the bounds by making assumptions about potential outcome

- Introduces a researcher degree of freedom that Chen & Roth (2024) show can sign-flip coefficients of interest

What Works: Normalized Estimands

'Percentage effects' need a base reference: percentage of *what*?

- One target parameter favored by Chen & Roth (2024) is *normalized estimands*, where linear estimands are re-scaled by some normalizing constant d
- d can be some functional of the outcome distribution (e.g., mean, median, SD) or an exogenous constant (e.g., cost of providing treatment)
- Normalized estimands can be interpreted in percentage units of d

Normalized estimands can be flexibly computed in most specifications as

$$\theta = \frac{\mathbb{E}\left[Y(1) - Y(0)\right]}{d}$$

and, for scale-equivariant estimators,

$$\theta = \mathbb{E}\left[\frac{Y(1)}{d} - \frac{Y(0)}{d}\right]$$

What Works: Poisson

For binary treatments, a natural choice for the normalizing constant is $\mathbb{E}[Y(0)]$

- Percentage effects in terms of $\mathbb{E}[Y(0)]$ can be consistently estimated using *Poisson regression* (Gourieroux, Monfort, & Trognon 1984; Santos Silva & Tenreyro 2006; Chen & Roth 2024)

A Poisson quasi-maximum likelihood model of the form $Y = \exp(\alpha + \beta_{\text{Pois}} X + \iota)$ can be estimated when $Y \geq 0$, and β_{Pois} is both scale-invariant and easily convertible into a percentage effect estimate:

$$e^{\beta_{\text{Pois}}} - 1 = \frac{\mathbb{E}\left[Y(1) - Y(0)\right]}{\mathbb{E}\left[Y(0)\right]}$$

However, Poisson models have a couple drawbacks

- **Con 1:** Can't be run on all data (requires no negative values and unique solutions)
- **Con 2:** Many useful estimation techniques do not have Poisson counterparts
- Due to these challenges, Poisson models could not be estimated for 45% of the papers in our replication sample

What Works: Quantile Regression

If the goal is to reduce the impact of outliers, use **quantile regression**

- **Pro 1:** Estimates a different parameter than OLS, which is far more robust to outliers
- **Pro 2:** Many empirical methods have quantile counterparts (e.g., Chernozhukov & Hansen 2005; Athey & Imbens 2006; Frandsen, Frolich & Melly 2012)
- **Con:** Sometimes infeasible due to computational cost

The right strategy for your project will depend on your data features

Conclusion

Our results imply that log-like specifications should be excluded from standard empirical practice

- It isn't just that log-like specifications fundamentally misidentify the (semi-)elasticities they are supposed to estimate (Cohn, Liu, & Wardlaw 2022; Chen & Roth 2024; Mullahy & Norton 2024; Thakral & Tô 2025)
- We show theoretically and empirically that by changing theoretically arbitrary parameters, log-like specifications can overfit the data and often generate spuriously significant results

Normalized estimands and Poisson regression remain useful for percentage effect estimation when there are zeros in the data, and quantile regression is useful for flattening outliers

- There is no one-size-fits-all approach; the optimal data analysis strategy will depend on your research question and data context

Our contribution: Ruling out specifications which we know can and do contribute to publication bias in the literature

References I

Aihounton, G. B. and A. Henningsen (2021).
Units of measurement and the inverse hyperbolic sine transformation.
The Econometrics Journal 24(2), 334–351.

Athey, S. and G. W. Imbens (2006).
Identification and inference in nonlinear difference-in-differences models.
Econometrica 74(2), 431–497.

Bellemare, M. F. (2018, Jun).
'Metrics Monday: Elasticities and the inverse hyperbolic sine transformation.

Bellemare, M. F. and C. J. Wichman (2019).
Elasticities and the inverse hyperbolic sine transformation.
Working paper.

References II

Bellemare, M. F. and C. J. Wichman (2020).
Elasticities and the inverse hyperbolic sine transformation.
Oxford Bulletin of Economics and Statistics 82(1), 50–61.

Brodeur, A., N. Cook, and A. Heyes (2020).
Methods matter: *p*-hacking and publication bias in causal analysis in economics.
American Economic Review 110(11), 3634–3660.

Burbidge, J. B., L. Magee, and A. L. Robb (1988).
Alternative transformations to handle extreme values of the dependent variable.
Journal of the American statistical Association 83(401), 123–127.

Chen, J. and J. Roth (2024).
Logs with zeros? Some problems and solutions.
The Quarterly Journal of Economics 139(2), 891–936.

References III

Chernozhukov, V. and C. Hansen (2005).
An IV model of quantile treatment effects.
Econometrica 73(1), 245–261.

Cinelli, C., A. Forney, and J. Pearl (2024).
A crash course in good and bad controls.
Sociological Methods & Research 53(3), 1071–1104.

Cohn, J. B., Z. Liu, and M. I. Wardlaw (2022).
Count (and count-like) data in finance.
Journal of Financial Economics 146(2), 529–551.

Daniele, G., M. Le Moglie, and F. Masera (2023).
Pains, guns and moves: The effect of the US opioid epidemic on Mexican migration.
Journal of Development Economics 160, 102983.

References IV

Frandsen, B. R., M. Frölich, and B. Melly (2012).
Quantile treatment effects in the regression discontinuity design.
Journal of Econometrics 168(2), 382–395.

Goldsmith-Pinkham, P. (2024).
Tracking the credibility revolution across fields.
arXiv preprint arXiv:2405.20604.

Gourieroux, C., A. Monfort, and A. Trognon (1984).
Pseudo maximum likelihood methods: Theory.
Econometrica 52(3), 681–700.

Lee, D. S. (2009).
Training, wages, and sample selection: Estimating sharp bounds on treatment effects.
Review of Economic Studies 76(3), 1071–1102.

References V

Mullahy, J. and E. C. Norton (2024).
Why transform Y? The pitfalls of transformed regressions with a mass at zero.
Oxford Bulletin of Economics and Statistics 86(2), 417–447.

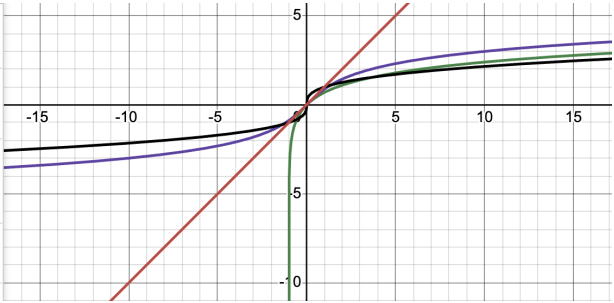
Norton, E. C. (2022).
The inverse hyperbolic sine transformation and retransformed marginal effects.
The Stata Journal 22(3), 702–712.

Santos Silva, J. M. and S. Tenreyro (2006).
The log of gravity.
The Review of Economics and Statistics 88(4), 641–658.

Thakral, N. and L. T. Tô (2025).
When are estimates independent of measurement units?
Boston University-Department of Economics.

Comparing the Functional Forms

1. $Y = \ln(X+1)$
2. $Y = \sinh^{-1}(X)$
3. $Y = X^{\left(\frac{1}{3}\right)}$
4. $Y = X$
5.



Back

Example Claim

To illustrate 'claims', consider an article in our replication sample (Daniele et al. 2023)

- The article's abstract makes two empirical claims defended by log-like specifications; we highlight these in blue and red and mark the location of their corresponding estimates [in brackets]

In this paper, we study how a positive economic shock to an illicit industry might foster migration. In 2010, a series of reforms to the U.S. health care system resulted in a shift in demand from legal opiates to heroin. This demand shock had considerable effects on Mexico, the main supplier of heroin consumed in the United States. We exploit variation in potential opium production at the municipal level in a difference-in-differences (DID) framework, which compares Mexican municipalities with different amounts of opium-suitable land before and after 2010. We find that people fled out of municipalities more suitable for opium production [Table 1], many to areas close to the U.S. border and into the United States. This is due to the increase in violence and conflicts [Table 6], as municipalities more suitable for opium became highly valuable to drug cartels. Overall, almost 95,000 people migrate within Mexico and 22,000 emigrate to the United States. Back

Within-sample Metrics

WithinSampleMetric$_{\Delta,s,a}$ can either be:

1. The absolute within-sample t-statistic
2. Dummy indicating within-sample statistical significance at the 5% level
3. Dummy indicating that the within-sample estimate sits in an empirical sweet spot
4. Dummy indicating that the within-sample estimate's t-statistic is outside the convex hull of the t-statistics for the within-sample linear and extensive margin specifications

Back to Overfitting

Parameterizing Robustness

Let $\hat{\beta}$ be the original regression coefficient of interest for estimate i and $\tilde{\beta}$ be the same coefficient from a given robustness specification s; we parameterize

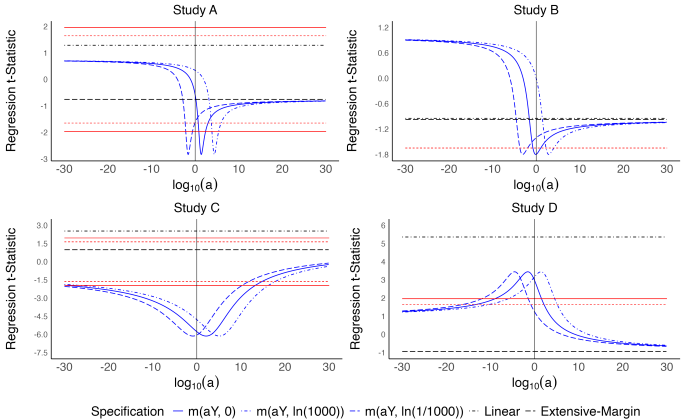
$$\text{Agree}_{i,s} = \begin{cases} \mathbb{1}\left[\tilde{\beta}\text{'s } (1-\alpha) \text{ CI excludes 0 and sign}\left(\tilde{\beta}\right) = \text{sign}\left(\hat{\beta}\right)\right] & \text{if } \hat{\beta}\text{'s } (1-\alpha) \text{ CI excludes 0} \\ \mathbb{1}\left[\tilde{\beta}\text{'s } (1-\alpha) \text{ CI contains 0}\right] & \text{if } \hat{\beta}\text{'s } (1-\alpha) \text{ CI contains 0} \end{cases}$$

$\text{Agree}_{i,s}$ thus indicates whether the robustness estimate's statistical significance conclusion agrees with that of the reproduction estimate

- As a second measure, we compute statistical significance dummy $\text{Sig}_{i,s}$ as an indicator for whether $\tilde{\beta}$'s $(1-\alpha)$ CI excludes zero

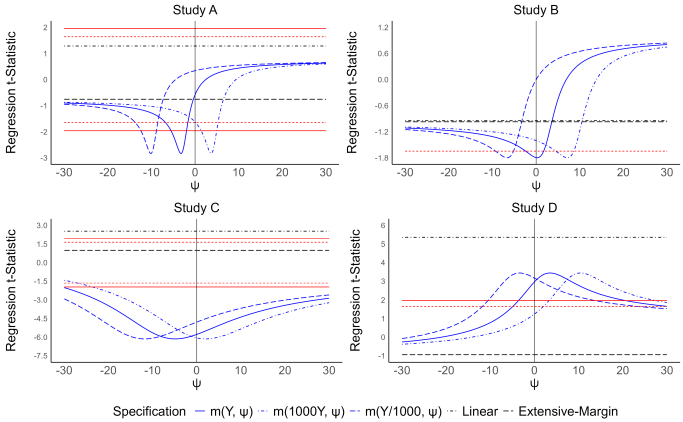
We parameterize $\text{Agree}_{i,s}$ and $\text{Sig}_{i,s}$ for $\alpha = 0.05$ and $\alpha = 0.1$ Back

Extensive Margin Calibration: Sweet spots in a



Back to Extensive Margin Calibration
Extensive Margin: Sweet spots in ψ

Extensive Margin Calibration: Sweet spots in ψ



Back to Extensive Margin Calibration
Extensive Margin: Sweet spots in *a*