Thread

Nick Hagerty
@hagertynw

2 useful papers today on how you can get basically any answer you want from regressions using arcsinh(y) or log(y+1)

I made these graphs to show the problem: The mean of the transformed var (black lines) varies like crazy by what units you choose for y

🧵 on my own take-aways

9:13 PM · Dec 12, 2022

93 Retweets **13** Quote Tweets **446** Likes

Tweet your reply

Reply

Nick Hagerty @hagertynw · 9h
Replying to @hagertynw
Here are the papers:

One by @jondr44 and @jiafengkevinc

> **Pedro H. C. Sant'Anna** @pedrohcgs · 17h
> Twitter is fun because we can promote the work of friends and people we admire!
>
> Today I want to call your attention to a brand new paper by @jondr44 and @jiafengkevinc, available at: jonathandroth.com/assets/files/L...
>
> What is the gist?
>
> Check the abstract!

1 4 18

Nick Hagerty @hagertynw · 9h

The other by @JohnMullahy and @healtheconnort1

> **Paul Hünermund** @PHuenermund · Dec 12
>
> New paper by @JohnMullahy and @healtheconnort1:
>
> "Why Transform Y? A Critical Assessment of Dependent-Variable Transformations in Regression Models for Skewed and Sometimes-Zero Outcomes"
>
> Show this thread

1 1 16

Nick Hagerty @hagertynw · 9h

Essentially: The 1 in log(y+1) is arbitrary (why not 10? 0.1? 0.0001?), and the arcsinh formula contains a similar hidden parameter

So they're sensitive to scaling, which isn't great because our answers should be the same whether we work with $, hundreds of $, millions, etc.

1 1 27

Nick Hagerty @hagertynw · 9h

(I think economists could benefit generally from thinking more about units! In physics EVERYTHING is about unit analysis, which gave me helpful intuition)

(I've been trying to tell people about this problem for years, though I've used the arcsinh too so I'm totally a hypocrite)

2 8 51

Nick Hagerty @hagertynw · 9h

Chen & Roth show that a "percentage" average treatment effect is *just not a well-defined estimand* when your outcome values can include 0.

I find this tremendously disappointing!! 😢 Though it seems obvious in retrospect, I think I'd been holding out hope for a way to get it

1 3 23

Nick Hagerty @hagertynw · 9h

Also: Mullahy & Norton show that when you place the 0's far from the rest of the distribution (when units are small or c in log(y+c) is large), you're basically just estimating a linear probability model!

Intuition: you're putting more weight on the extensive margin

3 3 20

Nick Hagerty @hagertynw · 9h

So what can we applied researchers do instead?

Here are the suggestions from both papers, along with my immediate reactions (which may or may not add value...)

1 8

Nick Hagerty @hagertynw · 9h

1. (Mullahy & Norton) Forget the proportional treatment effects, just estimate OLS on the untransformed outcome.

I don't like this because
(a) I often care more about the avg proportional change than the avg level change across individuals
(b) estimates can be noisy for skewed y

1 13

Nick Hagerty @hagertynw · 9h

2. (M&N) Use Poisson regression.

Poisson helps concern (b) but not (a). It directly estimates the log difference of the means in the treatment & control groups: log(E[Y(1)]) - log(E[Y(0)])

Not the average log difference: E[log(Y(1)) - log(Y(0))], which is what I tend to prefer

1 2 14

Nick Hagerty @hagertynw · 9h

Poisson gets at moments of the conditional distributions. It's not guaranteed to tell you about the avg proportional change across individuals

Just like how quantile treatment effects tell you how the quantile ITSELF moves, not the treatment effect for indivs at a given quantile

1 12

Nick Hagerty @hagertynw · 9h

In the framework of Chen & Roth, Poisson does not estimate a parameter of the form E[g(Y(1), Y(0)], at least not when you allow for arbitrarily heterogeneous treatment effects across individuals.

1 5

Nick Hagerty @hagertynw · 9h

3. (Chen & Roth) Estimate the treatment effect for a manually-calculated percentage change, like Y/Y_pre.

I like this idea, but:
- it requires panel data
- in my own experience it introduces a lot more noise
- it might not un-skew the variable if it isn't pretty autocorrelated

3 11