Notes of a Numerical Analyst

Straight Line Through Data

NICK TREFETHEN FRS

Recently, I had to find the linear slope of some noisy data, like this:

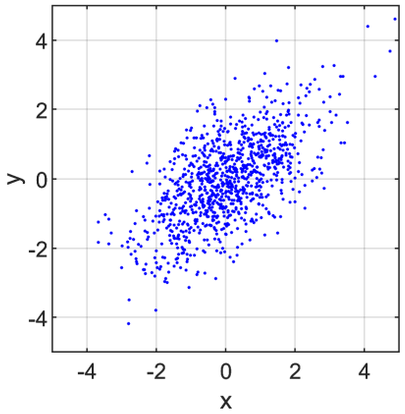

Figure 1. 1,000 data values, statistically symmetric in x and y

Of course, I did a least-squares fit, as we are all trained to do. The resulting slope was all wrong! I expected a line at $45°$, but this was much less than that. What's going on?

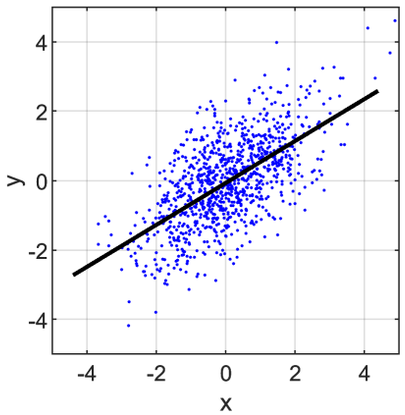

Figure 2. A least-squares fit gets the slope wrong

I showed a few mathematician friends, and they didn't see it. Of course, any statistician would know, so the first lesson of this episode is a reminder of the unfortunate separation of the mathematical and statistical communities. (Perhaps this is getting better in the new data science generation?)

The explanation is elementary. Ordinary least-squares fitting assumes that y depends on x, with noise just in the y direction. That's the asymmetry — and if you exchange variables and compute a least-squares fit of x to y, you get the slope anomaly in the other direction. (How did I reach this point of my career without knowing that?) If you switch to total least squares or principal component analysis (PCA), a properly balanced slope appears.

But digging deeper brings us into 140 years of the history of statistics [1]. It turns out that Fig. 2 encapsulates what Galton famously called *regression to the mean*. (This is where the term regression comes from.) In our dataset, x and y have the same distributions and a correlation (a term coined by Galton for just this discussion) between 0 and 1. If x is known to take a large value like $x = 4$, can we infer that y will probably also have a large value? Yes. Can we infer that its value will probably be as large as 4? No.

My application concerned weight vs height data and the BMI, the body mass index—a fraught topic going back to Adolphe Quetelet in 1832. If $w \approx Ch^{\alpha}$, what's the exponent α? Least-squares fitting of the log-log data I had on hand suggested $\alpha \approx 1.8$, but the plot looked even worse than in Fig. 2. Changing to PCA quintupled the slope to $\alpha \approx 8.6$! Such all-over-the-place results arising in such a simple problem remind us of a second lesson: to take *cum grano salis* the item in today's news cycle about the latest medical or sociological discovery based on a careful statistical analysis.

FURTHER READING

[1] S.M. Stigler, *The Seven Pillars of Statistical Wisdom*, Harvard, 2016.

Nick Trefethen

Trefethen is Professor of Applied Mathematics in Residence at Harvard University.