briefing note
313
March 2016

Big Data, Big Computers, Big Trouble

We are told that government, business, finance, medicine, law, and our daily lives are being revolutionized by a new found ability to sift through reams of data and discover the truth. But with big data comes big problems.

This article by Gary Smith was published on September 28, 2015 on Elsevier SciTech Connect.

Many years ago, James Tobin, a Nobel laureate in economics, wryly observed that the bad old days when researchers had to do calculations by hand were actually a blessing. In today's language, it was a feature, not a flaw. The calculations were so hard that people thought hard before they calculated. Today, with terabytes of data and lightning-fast computers, it is too easy to calculate first, think later.

Over and over, we are told that government, business, finance, medicine, law, and our daily lives are being revolutionized by a new found ability to sift through reams of data and discover the truth. We can make wise decisions because powerful computers have scrutinized the data and seen the light.

Maybe. Or maybe not.

RECOMMENDED READING

Artificial intelligence: Can Watson save IBM?
By Richard Waters

Published in The Big Read,
January 5, 2016

http://www.ft.com/cms/s/2/dced8150-b300-11e5-8358-9a82b43f6b2f.html#ixzz3wVfj95Wg

Two endemic problems are nicely summarized by the Texas Sharpshooter Fallacy. In one version, a self-proclaimed marksman completely covers a wall with targets, and then fires his gun. Inevitably, he hits a target, which he proudly displays without mentioning all the missed targets. Because he was certain to hit a target, the fact that he did so proves nothing at all. In research, this corresponds to testing hundreds of theories and reporting the most statistically significant result, without mentioning all the failed tests. This, too, proves nothing because one is certain to find a statistically persuasive result if one does enough tests.

In the second version of the sharpshooter fallacy, the hapless cowboy shoots a bullet at a blank wall. He then draws a bullseye around the bullet hole, which again prove nothing at all because there will always be a hole to draw a circle around. The research equivalent is to ransack data for a pattern and, after one is found, think up a theory.

A concise summary is the cynical comment of Ronald Coase, another economics Nobel laureate: "If you torture the data long enough, it will confess."

Do serious researchers really torture data? Far too often. It's how well-respected people came up with the now-discredited ideas that coffee causes pancreatic cancer and people can be healed by positive energy from self-proclaimed healers living thousands of miles away.

An example of the first sharpshooter fallacy is a study provocatively titled, "The Hound of the Baskervilles Effect," referring to Sir Arthur Conan Doyle's story in which Charles Baskerville dies of a heart attack while he is being pursued down a dark alley by a vicious dog:

The dog, incited by its master, sprang over the wicket-gate and pursued the unfortunate baronet, who fled screaming down the yew alley. In that gloomy tunnel it must indeed have been a dreadful sight to see that huge black creature, with its flaming jaws and blazing eyes, bounding after its victim. He fell dead at the end of the alley from heart disease and terror.

The study's author argued that Japanese and Chinese Americans are similarly susceptible to heart attacks on the fourth day of every month because in Japanese, Mandarin, and Cantonese, the pronunciation of four and death are very similar.

Four is an unlucky number for many Asian-Americans, but are they really so superstitious and fearful that the fourth day of the month—which, after all, happens every month—is as terrifying as being chased down a dark alley by a ferocious dog?

The Baskervilles study (isn't the BS acronym tempting?) examined California data for Japanese and Chinese Americans who died of coronary disease. Of those deaths that occurred on the third, fourth, and fifth days of the month, 33.9 percent were on day 4, which does not differ substantially or statistically from the expected 33.3 percent. So, how did the Baskervilles study come to the opposite conclusion? There are dozens of categories of heart disease and they only reported results for the five categories in which more than one-third of the deaths occurred on day 4. Unsurprisingly, attempts by other researchers to replicate their conclusion failed.

An example of the second sharpshooter fallacy is an investment strategy based on the gold-silver ratio (GSR), which is the ratio of the price of an ounce of gold to the price of an ounce of silver. Figure 1 shows that the GSR fluctuated around the range 34 to 38 during the years 1970 through 1985.

In 1986 one advisory service wrote that *The [GSR] has fluctuated widely just in the past seven or eight years, dipping as low as 19-to-1 in 1980 and soaring as high as 52-to-1 in 1982 and 55-to-1 in 1985. But, as you can also clearly see, it has always—ALWAYS—returned to the range between 34-to-1 and 38-to-1.*

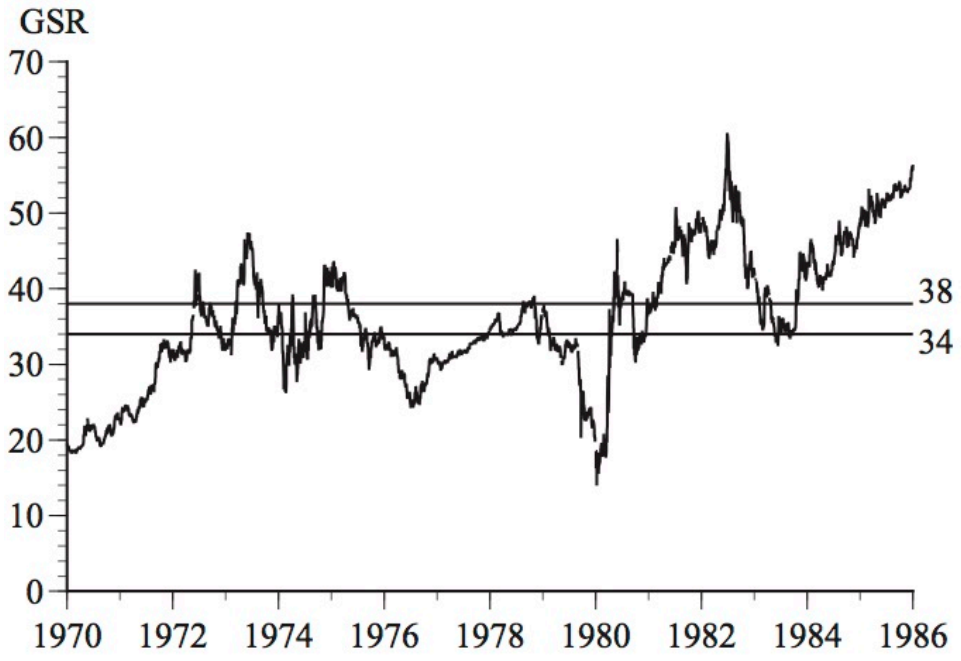

Figure 1 The GSR 1970–1985

The GSR strategy is to sell gold and buy silver when the GSR is unusually high and do the opposite when the GSR is unusually low. Using futures contracts to make these trade requires very little margin and creates enormous leverage and the potential for astonishing profits.

Like drawing the target after finding the bullet hole, discovering a pattern in ransacked data proves nothing more than that the data were ransacked for patterns. There is no logical reason why an ounce of gold should cost the same as 36 ounces of silver. As it turned out, after the GSR went above 38 in 1983, it did not come back until 2011. Leverage multiplies losses as well as gains and a 1983 bet on the GSR would have been disastrous.

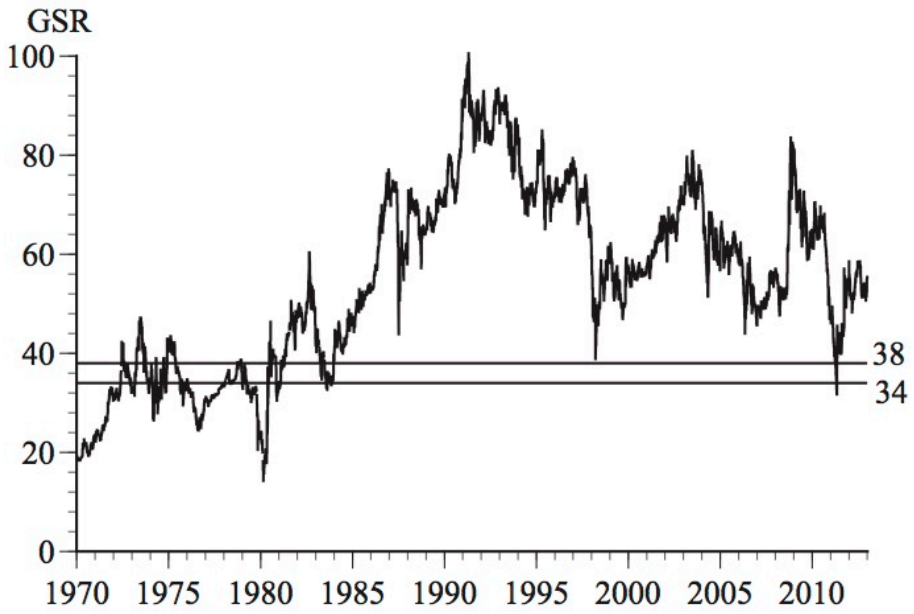

Figure 2 The GSR 1970–2012

The GSR is an early example of statistical arbitrage in which quantitative financial analysts ("quants") find a statistical pattern and assume that deviations from this pattern are temporary aberrations that can be exploited. If the ratio of gold to silver prices has historically been between 34 and 38, then, when the GSR goes outside this range, it must soon return.

Modern computers can ransack large data bases looking for much more subtle and complex patterns. But the problem is the same. If there is no underlying reason for the discovered pattern, there is no reason for deviations from the pattern to self-correct.

These two sharpshooter fallacies are examples of data mining. Thirty years ago, calling someone a "data miner" was an insult comparable to being accused of plagiarism. Today, people advertise themselves as data miners. This a flaw, not a feature. Big data and big computers make it easy to calculate before thinking, it is better to think hard before calculating.

About the Author

Gary Smith received his B.S. in Mathematics from Harvey Mudd College and his PhD in Economics from Yale University. He was an Assistant Professor of Economics at Yale University for seven years. He is currently the Fletcher Jones Professor of Economics at Pomona College. He has won two teaching awards and has written (or co-authored) seventy-five academic papers, eight college textbooks, and two trade books (most recently, Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie With Statistics, Overlook/Duckworth, 2014). His research has been featured in various media including the New York Times, Wall Street Journal, Motley Fool, NewsWeek and BusinessWeek. For more information visit www.garysmithn.com.

Gary's recently published book, *Essential Statistics, Regression, and Econometrics, 2nd Edition,* is innovative in its focus on preparing students for regression/econometrics, and in its extended emphasis on statistical reasoning, real data, pitfalls in data analysis, and modeling issues. This book is uncommonly approachable and easy to use, with extensive word problems that emphasize intuition and understanding.