BloombergView

ECONOMICS

Bad Data Can Make Us Smarter

OCT 29, 2014 11:53 AM EDT

By Noah Smith a A

At the Western Finance Association meeting this summer, I heard a presentation of an interesting paper titled "...and the Cross-Section of Expected Returns," by Campbell Harvey, Yan Liu and Heqing Zhu. The paper is sort of a finance version of the famous 2005 paper by John Ioannidis, "Why Most Published Research Findings are False."

The subject of both papers, in short, is data-mining. The number of published papers has exploded over the past century, but the statistical techniques used to judge the significance of a finding haven't evolved very much. The standard test of a scientific hypothesis is the so-called t-test. A t-test will give you a p-value, which is supposed to be the percent chance that the finding was random. So if you run a test and get a p-value of 0.04, many people will take that to mean that there is only a 4 percent chance that the finding was a fluke. Because 4 percent sounds like a low-ish number, most researchers would call such a finding "statistically significant."

Now, if there were only one scientific test of one hypothesis in all of human history, a p-value of 0.04 might be just as interesting as it looks. But in the real world, there are many many thousands of published p-values, meaning that a substantial number must be flukes. Worse, since only the tests with significant p-values tend to get published, there's a huge selection bias at work -- for every significant p-value you see in a paper, there were a bunch of tests that didn't yield an interesting-looking p-value, and hence weren't able to make it into published papers in the first place! This is known as publication bias. It means the publication system selects for false positives.

But it gets worse. Because the set of tests that researchers run isn't fixed -- since researchers need to publish papers -- they will keep running tests until they get some that look significant.

Suppose I ran 1,000 tests on 1,000 different totally wrong hypotheses. With computers, this is easy to do. Statistically, maybe about 50 of these will look significant with the traditional cutoff of 5 percent. I'll be able to publish the 50 false positives, but not the 950 correct negative results!

This is data-mining, and there's essentially no way to measure how much of it is really being done, for the very reason that researchers don't report most of their negative results. It isn't an ethics question -- most researchers probably don't even realize that they're doing this. After all, it's a very intuitive thing to do -- look around until you see something interesting, and report the interesting thing as soon as you see it.

As Harvey et al. show with a clever meta-analysis and simulation, this leads to scientific literature that's filled with false positives. They focus on factor models, which measure the tradeoff between risk and return. I'm sure that a similar result would be found if they ran their meta-analysis on back-tests of supposedly market-beating trading strategies.

It's a good paper, and a fun one as well. But should we be worried about this result? I can think of three reasons not to be too concerned.

First, Harvey et al. conclude that a more severe hurdle rate is necessary to take a test result seriously -- they suggest looking at results with t-ratios of 3, instead of the more commonly used value of 2 (a t-ratio is higher when the corresponding p-value is lower). But this might not be a long-term solution. A more severe cutoff might just increase the amount of data that researchers mine, and as data sets grow, we might soon be back to the same level of false positives. If the cutoff remains where it is, on the other hand, readers of papers can simply regard tests with t-ratios between 2 and 3 as being negative results. In other words, as long as people know not to take borderline "significant" cases too seriously, the current system allows the publication of a bunch of what are effectively negative results, which is a good thing.

Second, the idea that false positives are false research findings takes a pretty severe view of how science works. Although the popular press often hypes results based on one eye-catching paper, scientists themselves tend to be much more circumspect. A "positive" result is really just a preliminary, exploratory finding. It's a way of telling scientists where to look in the future. As more scientists follow up on a research finding, they either replicate it, or find that it can't be replicated. Science is a slow, iterative process.

In the life sciences such as biology and neuroscience, replication tends to be very expensive, which is a problem. But in finance, all you really have to do is to wait for more data to come in,

and test models again with the new data. Good, solid findings will endure, while the results of data-mining will mostly vanish.

In other words, published results shouldn't be regarded as true or false. They are just steps along the road to greater understanding of the universe around us. For those who are used to demanding immediate, definitive answers, the glacial progress of science may seem maddening. But if what you want is the truth, you've got to be patient.

To contact the author of this article: Noah Smith at noahsmith.bloomberg@gmail.com.

To contact the editor responsible for this article: James Greiff at jgreiff@bloomberg.net.