portfolio construction forum

The trend that is ruining finance research

Michael Edesess | EDHEC-Risk Institute | 07 September 2017

According to Andrew Ang, a guru of factor-based investing and former chair of the finance and economics division of Columbia Business School's Data Science Institute, the "anomalies" literature is the scientific foundation for quantitative asset management. But this focus, which was not very scientific to begin with, is proving its utter ruin.

Anomalies are instances of investment strategies or subgroups of securities that have outperformed the market on a risk-adjusted basis over a long period of time. They are called anomalies because they stand in contradiction to efficient market theory, which would imply that such phenomena cannot occur. The protocol for testing for the existence of an anomaly has been established for some time, at least since two papers in 1992 and 1993 authored by Eugene Fama and Kenneth French. Those papers explored what appeared to be the value stock and small-stock anomalies. The testing protocol involves regression and hypothesis testing to determine if the anomaly is statistically significant.

I will first review the subject of hypothesis testing, then discuss three recent papers that show how badly off-track most anomalies research is really.

R.A. FISHER'S WORK ON THE FARM

Perhaps it was because Ronald Aylmer Fisher had been forced in his youth, for lack of funds after graduating from Cambridge University, to take a job working on a farm in Canada. Or, perhaps it was simply because agriculture was much more of a vibrant developing industry in 1919 than it is now. Or, perhaps it was because Fisher had gotten into a professional scrape with his mentor, the famous statistician Karl Pearson. But whatever the reason, when in that year – 1919 – Pearson offered Fisher a plum university job, Fisher turned it down and took a job instead at the Rothamsted Agricultural Experiment Station, in the English county of Hertfordshire about 50 kilometers north of London.

This turned out to be a boon to statistics. At Rothamsted, Fisher pioneered two of the field's most important breakthroughs, hypothesis testing and the design of experiments. These tools are now used in the vast majority of scientific studies of empirical data.

Fisher's need was to test whether grain variety A, a new grain he was testing, would produce a greater yield than the old grain variety B. So he planted a number of plots of land with variety A and the same number with variety B – taking care that there were no other

systematically different factors affecting the plots planted with A and those with B, such as sunshine and shading.

One of the grains, say A, would inevitably produce a higher average yield than the other, since it was unlikely that the average yields would be exactly the same. The challenge was to decide whether the difference was large enough to conclude that A produces higher yield than B.

Fisher's breakthrough was to pose this question: If there really wasn't any difference between A and B (the "null hypothesis") how probable is it that, in the experiment, A's average yield would be as much greater than B's as it was?

The probability that answers that question is called the p-value. (Fisher calculated p-values with the help of a researcher called "Student" who was actually William Sealy Gosset and worked for the Guinness brewing company in Ireland. Gosset used the pseudonym Student because Guinness didn't allow its researchers to publish, for fear they might reveal corporate secrets.) Conventionally, if the p-value is less than 0.05 (one chance in 20) it is concluded that the result of the experiment is unlikely to have occurred by chance; therefore, A must produce higher yield than B.

THE METHOD DOES NOT SCALE

Fisher's method has become the most-used technique in statistics. But it took more than 80 years for researchers to realise that the methodology does not scale. When it is applied a large number of times, its validity breaks down. And it is indeed used many times – tens or hundreds of thousands times a year, or more – by researchers in a wide variety of fields.

When you reject a null hypothesis because, if it were true, the results of your experiment would have occurred only once-in-20 trials, this also means that for every 20 experiments you do, you will come to a wrong conclusion. Then what if you do 1,000 experiments – not to mention tens of thousands? If Fisher had tested 1,000 grain varieties, he would have concluded that at least 50 of them – one in 20 – produced higher yield than the old grain.

Suppose that Fisher had done this, and then reported that he had discovered two or three miraculously improved grain varieties – and proceeded to profit from them – but didn't tell anyone that he had actually researched 1,000 of them, and the vast majority didn't pass the significance test. And now suppose that many years later, a historical researcher discovered this fact about Fisher. And suppose also, that the grain varieties he identified as miraculous were later shown not to have outperformed at all. Fisher's name would be besmirched, and his status in the history of statistical innovation would drop many pegs.

But financial researchers do not suffer such a comeuppance for the same subterfuge.

CAMPBELL HARVEY'S PRESIDENTIAL ADDRESS TO THE AMERICAN FINANCE ASSOCIATION

As long as Fisher did a limited number of experiments, his method worked well enough. But with the large number of hypothesis tests being applied now, in many fields, it doesn't work well. This was finally pointed out forcefully by medical researcher John Ioannidis in a landmark paper in 2005 titled, "Why most published research findings are false".

Anomalies research in finance is especially cursed with the scaling problem, not least because there is so much financial data available – most of it too easy to access and manipulate – but also because the pressure to publish or to get marketable results has caused researchers to leap into data studies without first doing careful reasoning.

When Campbell Harvey gave his presidential address as the newly-elected president of the American Finance Association in January, he made this the subject of his talk. He elaborated on the system that increases the status of academic journals – and of the academics who publish in them – based on the number of citations of their publications.

Papers that obtain statistically insignificant results are less likely to be cited, therefore they are much less likely to be published or even to be submitted for publication.

Hence – like my hypothetical version of R. A. Fisher who didn't tell anyone that he had done 1,000 experiments before reporting the "significant" results of a few of them – aspiring submitters of papers for academic journals submit only their papers that report significant results, and do not publicise their studies, of which there may be many, that did not obtain significant results. In consequence, like the medical research papers that Ioannides identified, and as Professor Harvey himself has suggested, most published research findings in finance are false.

MORE PROBLEMS WITH ANOMALIES RESEARCH

But as three recent papers point out, p-hacking (the name for doing too many studies in order to get one nominally significant p-statistic) is only the beginning of anomalies research problems.

These problems are identified in a paper by Jae H. Kim titled "Stock Returns and Investors' Mood: Good Day Sunshine or Spurious Correlation?" and its predecessor, co-authored with Philip Inyeob Ji, "Significance testing in empirical finance: A critical review and assessment," and in a paper by Kewei Hou, Chen Xue, and Lu Zhang titled, "Replicating Anomalies".

In addition to "widespread p-hacking" the authors identify three other sources of errors in anomaly studies:

1. Use of linear regression, the result of which is unduly influenced by outlying data;

2. Techniques that tend to equal-weight stocks, resulting in findings that are dominated by microcaps; and,
3. Keeping the p-value threshold constant, whether at 0.05 or another value.

The third source of error is the hardest to explain, but the most illuminating, so I'll save it for last.

Methods that overweight the influence of small stocks

Both sources of error (1) and (2) above tend to overemphasise the influence of small stocks, which are only a small part of the market.

Picture an x-y plot with a cloud of points, and imagine drawing a straight regression line through the cloud in order to approximate its trend. The regression formula calls for drawing the straight line by minimising the sum of the squares of the deviations of the points from the line – so the points that are farthest from the line have an exaggerated influence on where the line is placed.

This means that more volatile stocks, which are more likely to be outliers, will have more influence on the resulting slope of the line (its beta). I have also shown that the effect is exacerbated by the common – and erroneous – practice of using ordinary monthly holding-period returns in the regressions instead of logarithmic returns. (For example, Fama and French in their 1992 and 1993 articles employ holding-period returns, not logarithmic returns.)

This means that small stocks, which tend to have higher volatility, will have a greater influence on the result than large stocks. Hence, because anomalies are more likely to exist for less-intensively-analysed small stocks than for well-followed large stocks, the conclusion that an anomaly exists in the market as a whole may be mistakenly drawn when it may only exist for very small stocks).

In addition, techniques that tend to equal-weight stocks result in findings that are dominated by microcaps. Microcaps comprise 60% of stocks but only 3% of market value. Price data for microcaps is often less reliable – and more volatile – than for other stocks. Therefore, like the use of linear regression, techniques of analysis that tend to treat stocks equally are prone to conclude that there is an anomaly in the market as a whole, when that anomaly may exist only for microcaps.

Keeping the p-value threshold constant

This phenomenon is prominent in the papers authored and coauthored by Jae Kim – but, as Kim notes, it has also been noted by such leading lights as Nobel Prize-winner Kenneth Arrow.

Let's use an example to show the hazards of hypothesis testing using a fixed p-value, such as 0.05, as the threshold for significance, by posing the question: How likely is it that when your experiment concludes that it is an anomaly, it really is an anomaly? This is not the same as the question as to whether or not to reject the null hypothesis.

Our example will answer the former question for a specific set of numbers. The results may be surprising.

Suppose that a particular anomaly is such that the probability of correctly detecting it is 90%. The probabilities of the four possible experimental outcomes are as shown in Figure 1.

Figure 1: Probabilities of the four possible results of an anomaly test

		Probability of concluding that:	
		It is an anomaly	It isn't an anomaly
Given that:	It is an anomaly	90%	10%
	It isn't an anomaly	5%	95%

Source: Michael Edesess

The 5% in the lower left is the p-value threshold – the cutoff probability at which it is decided that the null hypothesis, "not an anomaly", is rejected.

Now let's put numbers in. Suppose 500 anomalies are tested for, and suppose that in 4% of the cases (20 of them) the phenomenon being tested for actually is an anomaly. The numbers of tests in each category are shown in Figure 2.

Figure 2: Results of 500 anomaly tests

		Conclusion is that:		
		It is an anomaly	It isn't an anomaly	Total
Given that:	It is an anomaly	18	2	20
	It isn't an anomaly	24	456	480

Source: Michael Edesess

Notice that if the test concludes that an anomaly exists (the "It is an anomaly" column), the odds are 24 to 18 that it actually isn't an anomaly. This result is not obvious from the 0.05 cutoff for the p-value in the significance test. In short, if your test concludes that an anomaly exists, the chances are greater than 50-50 that you're wrong.

All the researchers whose papers are cited here, including Harvey, advocate using a Bayesian approach instead of Fisher's hypothesis testing approach. The Bayesian approach allows you to say what the probability is that you've found an anomaly – something that the hypothesis testing approach doesn't allow you to do, though many people mistakenly believe it does. But the Bayesian approach has had difficulty taking hold, in large part because it first requires a "subjective" estimate of the probabilities by the researchers. The estimate is then amended as a result of the experiment. But it is still seen as a contamination of research objectivity.

THE REAL BOTTOM LINE

It's awfully hard to find it in any of the papers, but it's there, just once, in the paper by Kim and Ji – they cite a paper by E. Soyer and R. M Hogarth titled, "The illusion of predictability: how regression statistics mislead experts." According to Kim and Ji, the authors of that paper report that "regression statistics and statistical significance create an illusion of predictability: their survey reveals that economists provide better predictions when they are presented with a simple visual representation of the data than when they make predictions purely based on statistical measures such as R-squared or t-statistic."

Isn't this evidence enough? Would it be possible to cram the over-quantification and over-mathematisation demon back into the bottle? Can't we see that an insistence that academic papers have to have a quantitative mathematical analysis, usually involving a routine and

tiresome application of regression, is causing a form of lobotomisation of the research brain?

It bears repeating: "regression statistics and statistical significance create an illusion of predictability". And the authors even give a viable alternative: "economists provide better predictions when they are presented with a simple visual representation of the data than when they make predictions purely based on statistical measures."

In every one of these papers, this is what the conclusion should be. The pressure to regress is overwhelming; it's past time for an organised resistance.

Michael Edesess is adjunct associate professor and visiting faculty at the Hong Kong University of Science and Technology, chief investment strategist of Compendium Finance, adviser to mobile financial planning software company Plynty, and a research associate of the Edhec-Risk Institute.

This article is abridged and reproduced with permission from Advisor Perspectives.