Epistemology and the Incentives of Investment Managers

**Introduction**

Epistemology, i.e. how we know what we know, represents today's largest challenge in quantitative finance. The recent explosion in data and computing power (not to mention institutional incentives) induced the irrelevance of most classical tools of inference allowing researchers to use them to 'prove' pretty much whatever they want. A recent paper titled [Replicating Anomalies](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2961979) by Hou, Xue, and Zhang tested 447 anomalies in the finance literature, successfully replicating only 35%. In the following example I'll try to demonstrate the concept of '[Researcher Degrees of Freedom](http://andrewgelman.com/2012/11/01/researcher-degrees-of-freedom/)', the ability of researchers to make experimental design choices until they get the results they are looking for.

Recently I read a [finance Twitter conversation](https://twitter.com/KeyserSozeCFA/status/847535182467018753) discussing the incentives of mutual fund managers. The writer discussed working at a mutual fund who killed it (that's a technical term by the way) over the previous year, outperforming its benchmark by 10%. As a young analyst he was excited for the following year as continued out-performance would propel them through the mutual fund ranks. To his dismay, the portfolio managers decided to reduce their active risk as the fund track record was now good enough to raise substantial assets with and the cost of under-performing and spoiling it was higher than the benefit of improving it. This behavior (lowering active risk after periods of substantial outperformance) fits with the claim that many mutual funds behave more as "Asset Gatherers" than "Asset Managers" and so if it were true you would expect it to be pervasive. I got my hands on a data set of US mutual fund returns and as any good empiricist put this claim to the test.

**Design**

This is where the 'Researcher Degrees of Freedom' get involved. In order to test this hypothesis I could have chosen any of a series of experimental designs. The first decision is the test variable. Given my data I decided on realized tracking error as my measure of 'active risk'. Ideally, with access to holdings, my proxy of active risk would either be Active Share or the forecast tracking error from a risk model like Barra as holdings based measures are less noisy. Investors, obviously, don't know what their realized tracking error will be and many instead rely on risk models to position their portfolios given their risk preference. So, if for a specific portfolio we saw a large change in forecast tracking error we would be confident this was something intended by the manager instead of the vagaries of market volatility. At this point we have three degrees of freedom (realized tracking error, forecast tracking error, active share).

Having decided on the test measure the next step was determining a cutoff between what we would expect *a priori* to be high and low active periods. First, we need to decide on the outperformance threshold and period. In other words, over what period and given what amount of outperformance would we expect a manager to decrease his active risk. I decided to use 10% outperformance over the previous twelve months. I chose these numbers because they were what was mentioned in the Twitter conversation and also were round numbers that seemed reasonable, moreover prospective investors tend to look at trailing one year returns before making an investment. A researcher looking to 'p-hack' his way to statistically significant results would likely try several combinations of these numbers until he found results he liked. In terms of outperformance seemingly reasonable values are 5%, 10%, 15%; or he could also look at percentiles of outperformance (rank relative to universe) and look at say 90%, 95%, 99%. In terms of period, he could try 6 months, 12 months, 24 months all while seeming reasonable (if you would like to try your hand at this p-hacking here's an interactive tool from [fivethiryeight](https://fivethirtyeight.com/features/science-isnt-broken/#part1)). The next step is deciding the length of the period to calculate tracking error over both before and after the cutoff. Again, an unscrupulous researcher could try any series of combinations until he stumbled upon statistically significant results. I decided to test 6, 12, and 24 months before and after the cutoff. At this point we are up to 162 different combinations a researcher could try (3 variables \* 6 return thresholds \* 3 outperformance periods \* 3 tracking error period = 162), making it much easier to find statistically significant results.

A good trick if you run into a situation like this is to use the [Bonferroni Correction](https://en.wikipedia.org/wiki/Bonferroni_correction) to test the robustness of the results. This correction tries to counteract the problem of testing many hypothesis. It simply states that for a desired confidence level (say 5%), and given multiple tests performed (say 10), or that you suspect the researcher could have performed, you divide your confidence level by the number of tests (5%/10 = 0.5%).  This corrected confidence is the p-value each test should have to obtain significance at your original level.

To summarize, the process is the following:

* For each manager (n = 1721) calculate rolling 12 month cumulative alpha (for the period June 2006 - June 2016)
* For each month where the rolling alpha exceeds 10% calculate the tracking error of the previous and following m months (6, 12 or 24)
* Run a paired T-Test of the before and after with the alternative hypothesis being that the before tracking error is greater than the after

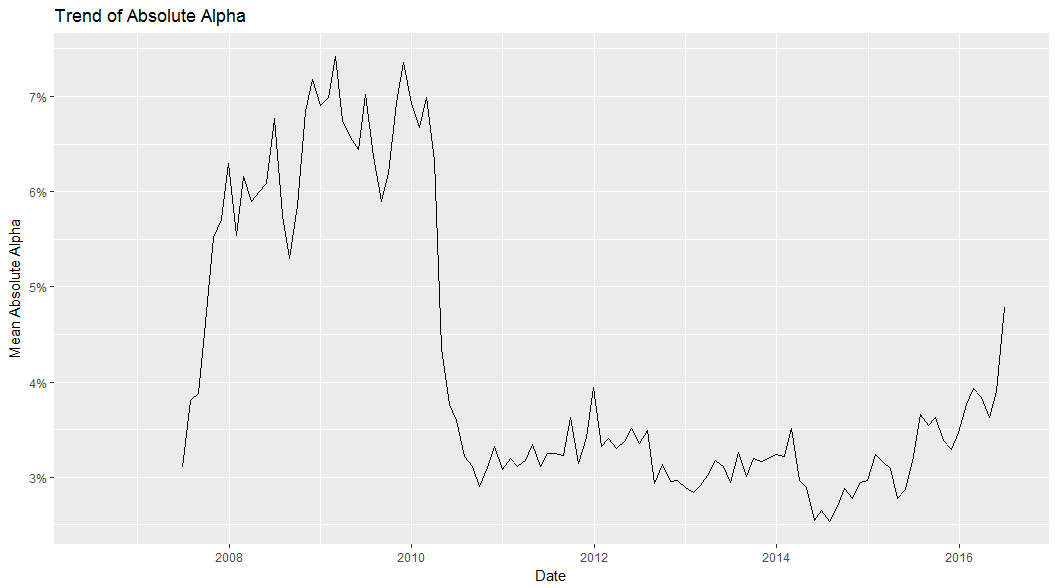
For anyone interested in replicating my analysis (performed using R) all my code and data can be found at the following [github repo](https://github.com/rjvelasquezm/Manager_Incentives).

**Analysis**

The results (table below) support our initial hypothesis with the mean difference ranging between 9 bps and 21 bps. It also raises an interesting observation, the shorter the time period the smaller the effect.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Before TE | After TE | Mean Difference | Mean Difference (percentage) | T-Statistic |
| 6 Months | 2.12% | 2.03% | -0.09% | -4.2% | 7.23 |
| 12 Months | 2.2% | 2.05% | -0.14% | -6.4% | 10.46 |
| 24 Months | 2.14% | 1.93% | -0.21% | -9.8% | 17.8 |

There's a few ways we could potentially explain this. One, is that changing portfolios takes time so the shorter time periods have not been able to fully capture the manager's intention which is more apparent over longer time periods. This explanation fits well within our framework. An, alternative explanation though could be that over time the tracking error of managers has tended to decrease. If this is true our results might be more a product of this trend than our hypothesis. The below chart shows the mean absolute alpha of all managers by month. It confirms our alternative explanation as tracking error has indeed dropped through time. This reduces the confidence that our results are actually due to our theory.



Taking this information into account, an alternative test could be to look at a manager's tracking error as compared to all other managers instead of the absolute number. In other words does their rank in terms of tracking error decrease after periods of high performance (say moving from the 75% percentile to the 70%). This allows us to continue working with tracking error but also lets us sidestep the issue of the long term trend. The below results are encouraging and support the explanation that the time trend (6 vs. 12 vs. 24 months) in differences is a result of the time it takes managers to reposition their portfolios.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Before TE Percentile | After TE Percentile | Mean Difference | T-Statistic |
| 6 Months | 65% | 63.6% | -1.3% | 5.01 |
| 12 Months | 67.6% | 65.4% | -2.2% | 9.67 |
| 24 Months | 65.8% | 63.1% | -2.6% | 11.45 |

At this point, a reader might reasonably conclude that the researcher has done a good job of testing his hypothesis and looked at alternate possibilities. Particularly, we've tried to disprove our results on two dimensions, length of test period, and relative to the long term trend. Given this, he might be inclined to accept the researcher's hypothesis that mutual fund managers tend to decrease their active risk after periods of outperformance. That is of course, before I tell you about the test that did fail.

As a final robustness test I decided to exclude the period over which the alpha was calculated in computing the 'Before TE'. The logic for this is simple; mathematically we would expect tracking error to be higher in periods of high outperformance than in periods of low outperformance, somewhat independently of the manager's actual active risk intention (if we had holdings based ex-ante tracking error this would not be an issue).

The below results disprove our hypothesis. Over all periods the difference ranges positive 21-26 bps which is larger both in absolute and percentage terms to the difference we were seeing when we included the outperformance period in the calculations. If you believe this to be a better measure of 'manager intention' then it seems managers actually increase their active risk after periods of outperformance. Some post-hoc rationalizations for this are: managers become more confident in their skills, managers judge the opportunity set to be better, the outperformance skews their portfolio weights causing an indirect increase in active risk.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Before TE | After TE | Mean Difference | Mean Difference (percentage) | T-Statistic |
| 6 Months | 1.82% | 2.03% | 0.21% | 11.5% | 12.15 |
| 12 Months | 1.79% | 2.05% | 0.26% | 14.5% | 17.33 |
| 24 Months | 1.71% | 1.93% | 0.22% | 12.9% | 18.06 |

**Conclusion**

As we have seen it is very easy for researchers to produce results that fit their hypothesis (not to mention biases) and report only on the tests performed that supported the hypothesis while failing to mention those that did not. This is the problem of what Nassim Taleb refers to as 'Silent Evidence'. In research and many other situations in life we only see what succeeded, not what failed. It may be that someone performed the same research I am reporting on using a different data set or technique and found no evidence of tracking error reduction and put his results in a drawer, never to be seen again. After all, research journals and careers are made out of positive results not negative ones.

As to the question of whether managers actually behave in this self-interested way, I think the results of this analysis are inconclusive. Given my knowledge of the investment industry I had a moderately strong prior that managers would engage in this type of behavior. The results have left me with about the same confidence I had before. I'm pretty sure some managers are engaging in this behavior but it may not be as pervasive as I thought. The main takeaway is to always question study conclusions, especially novel ones. Ideally, to accept a hypothesis you should have what Ray Dalio calls 'Triangulation', which are results that confirm the hypothesis through different unrelated ways. You would want to see several studies that test the hypothesis using different methods by different researchers with different datasets and in different markets. A good example of this is the momentum anomaly, which shows up virtually everywhere it has been tested (yes, including [Japan](https://www.aqr.com/library/journal-articles/momentum-in-japan-the-exception-that-proves-the-rule)), and oh by the way, apparently is present in manager outperformance as well (chart below).

If you liked my writing and ideas you can find more of it on my [Medium site](mailto:Medium%20site) and feel free to reach out at rjvelasquezm92@gmail.com.

