EVALUATING QUANTITATIVE INVESTMENT STRATEGIES

A practical guide to determining how good that back-test really is

Maria Schiopu, Milliman Financial Risk Management LLC

Introduction

In today's investing environment sophisticated strategy formulations powered by novel computing technologies, reliant on complex data and advanced calculations are merely mainstream fare for savvy investors. But as the industry has leapt from fundamental investing to complex mathematics (see SSGA paper), enormous pools of assets remain under the guardianship of individuals who are not always fluent in financial engineering and its many dialects; although they are of course vastly experienced in finance and investing. Nonetheless, such individuals serving on investment committees and boards of trustees for pensions, endowments, variable insurance trusts, etc. have a responsibility to expand their skillset to meet this challenge. In particular, as the insurance industry globally has been moving away from asset intensive general account products, and favoring separate account products, boards of trustees assume much greater responsibility for prudent oversight of investment strategies. This paper is intended to help them learn to speak the language of quantitative investing and build a toolkit to evaluate complex strategies.

Background

A simple Google search will quickly reveal a glut of online materials on evaluating quantitative investing strategies. Unfortunately, these resources appear to generally fall into one of two extremes: either too technical, or entirely qualitative, rendering them of limited practical use.

Academic papers, although thorough and interesting, are clearly written by quants for quants, and are sure to leave many non-quant investment committees quickly remembering the type of math-anxiety they left behind in high school. For those looking for such entertainment, a few examples are available here, here, here, and here. On the other end of the spectrum, are myriad of non-technical articles and blog posts on the topic, which seem to do little beyond point out pitfalls and dole out dire warnings, without offering any practical evaluation methods.

For the remainder of this text I will assume that you, the board, have already outlined a clear investment objective, and that a prospective investment manager has presented you with a strategy proposal. As is often the case with such proposals, the manager has provided a descriptive presentation of the investing algorithm, perhaps various pieces of evidence to support the strategy rationale, and a back-test of the strategy.

Before we dive any deeper, a back-test consists of a historical hypothetical calculation of the proposed strategy's results. In essence, the manager is attempting to demonstrate, to the best of their ability, how your investment would have performed if you had already been implementing their strategy for some time in the past. As it answers this important question, back-testing is arguably the single most widely used tool of strategy evaluation. But beyond evaluation, it has arguably become a standard method of strategy formulation as well, as data mining has increasingly made it easy for managers to skip hypothesis formulation altogether, and dig right in. In the introductory remarks of their paper on machine learning, Arnott, Harvey, and Markowitz give an excellent explanation of how this came to be. Although the authors do not endorse the use of data mining as the primary tool for strategy design, and neither do I, its widespread use raises the importance of developing an effective evaluation framework.

Back to our hypothetical board members, you might be tempted to think that at this stage, if the investment algorithm and rationale make sense, and the back-test meets your performance goals, then your job here is done. Nonetheless, the challenge lies in the fact that the back-test and strategy design itself, by virtue of being developed after looking at the underlying market data, undeniably benefit to some extent from hindsight. Moreover, the observed market data represents only a sample of the true possibilities, and it is not clear how representative this sample really is of future market dynamics. After all, as the old adage goes, past performance is not indicative of future results.

In other words, the back-test you are looking at is guaranteed to be “overfitted” to the historical data sample that the strategy was designed on (instead of being fitted to the true population of future market data, which is of course unobservable). The question isn’t whether or not overfitting is present, but rather to what extent is overfitting likely to lead to a deterioration of results once the strategy faces live implementation in the markets. It is generally the case that the better a back-test looks, the more overfitted it is. So, when it comes to back-tests, the good ones actually are likely better in practice than the best! Read on to discover some common-sense practical methods you can use to zero in on a “just good enough” investing strategy, that is more likely to be a solid performer in practice. We will focus on three key areas of inquiry, complete with a neat sample checklist of key questions, and some practical examples to illustrate how they can help root out unsuitable, erroneous, or impractical strategy designs.

Does the Strategy Suit Your Goals?

You might be tempted to try to rephrase the question as “Does the back-test exceed the desired return?”, but that would be underestimating the seriousness and complexity of this question. A more realistic rephrasing would be something like: “Does the back-test exceed the desired long-term return, while also exhibiting sufficiently reliable short-term performance?” The reason for the latter addition is that, although intuitively we all understand that any investment is bound to have some periods of underperformance, we are much more likely in practice to change strategies when that happens. However, the key to reaching those long-term average returns is to stick with it, and resist the urge to tweak or abandon the strategy along the way.

Another aspect worth noting is that risk objectives are just as important as return objectives, although we won’t go into the details of setting appropriate strategy goals. The obvious example is investing in the context of insurance, especially in the increasingly lengthy global low interest rate environment. It is imperative for insurance company boards to understand the finer points of monitoring and evaluating strategies in the context of providing both growth and protection to their policyholders.

So in essence, think of strategy goals as two-fold: first, meeting your traditional investment return and risk goals, and second, passing your periodic reviews. For that purpose, make a plan from the start for how you will evaluate ongoing investment performance. If you plan on reviewing results on a quarterly schedule, for example, ask the investment manager to provide back-tested partitions based on that frequency. Then approach those results with a realistic mindset and test your reactions to each quarter’s results, as you would with live performance, based on a trusted set of performance statistics. What proportion of those quarterly results meet your goals? Are there stretches of underperforming quarters that lead you to believe you would have dropped or altered the strategy?

As an example, let’s say you’re investing in the 500 largest companies by market capitalization in the US, and your goal is to make 8% returns per year. Well, if you had followed the S&P 500 Index from 1990 to the end of the most recent quarter (3/31/2020), you’d be happy to learn that on average you’ve achieved just over 9% annualized returns, so you have met your goal. But if you had planned to evaluate performance on a quarterly basis, then you should know that only 73 of the 121 quarters had returns exceeding 2%. Furthermore, during the global financial crisis, the investment would have gone six straight quarters without a positive return. What’s more, if you had implemented the strategy at the start of 1990, you would have found that in your first five years, almost half the time the strategy results were disappointing (9 out of 20 quarters from

1990 to 1994 had returns of less than 2%). You would have still met your annual return goal of 8%, but certainly not in a linear way.

To reiterate, the point is that for a strategy to work for your goals, it has to actually work for your review schedule, because otherwise you're likely to drop it or alter it. Additionally, do not underestimate the extraordinary power of hindsight, that lets you benefit from knowing already that the S&P 500 proved to be a great investment in the '90s. When you are evaluating live performance, the temptation to drop or tweak an underperforming strategy is very real.

Sample Questions on Suitability

- Does a descriptive outline of the investing algorithm reasonably match your investing purpose?
- What are the expected performance statistics over various time frames?
- What scenarios are likely to lead to over- and underperformance?
- Are underperforming scenarios correlated with other business factors that affect your institution?
- What allocation limits or restrictions are applicable to this mandate, and does the proposed strategy respect them?

Was the Back-test Well Constructed?

The topic of good and bad practices in back-testing is vast enough to be a common subject among PhD theses, but I will focus on a few common faults of sloppy back-test design and give you some methods to effectively spot these red flags. This is the trickiest and most technical part of your task, but the good news is that throughout this difficult examination, you will also learn a lot from the manager's approach to your queries, as well as develop a much stronger and more open partnership. The two main topics that must be tackled here are data and algorithm.

First, in regard to the data, your main goal is to determine whether a credible data sample was used to construct the back-test. Ask the manager to explain how they selected the data sample, and whether the complete historical data was utilized in the construction of the back-test. If parts of the available data were excluded, then see if they provide sensible reasoning for doing so.

Much ado has been made about the topic of partitioning data into training versus testing samples in relation to back-testing. I argue that this method is completely irrelevant and unsuitable to investment back-testing. Although extremely useful in many data-driven endeavors, this approach requires large data sets that can be sectioned randomly to create similar, but independent samples.

Something like car insurance claims data within a state can be sectioned this way by randomly picking 20% of policyholders across each county in the state, for example. However, it is obvious that picking all policyholders from the top largest counties until we reach 20% of the state's population, has a slim chance of producing representative training and testing samples. By a similar logic, training an investing algorithm on data from 2000 to 2015, and then testing its performance from 2016 to present bears little value if any, as the two periods are not independent or similar. In particular, pay attention to managers that do claim to have performed "out-of-sample testing" and the types of claims that they make about this testing.

So, training versus testing data segregation does not make sense for investment back-testing, but it does bring up an important point: how much historical data should be used? Market dynamics and structure (e.g. impact of high frequency trading on electronic exchanges) have changed over time, and so recent years' data is most relevant. At the same time, economic cycles unfold on the scale of decades, so recent data that most accurately represents market structure may not include enough variety. Unfortunately, there is no right answer to this question, but there are many wrong ones. A good manager will demonstrate that they have thought about this deeply and critically and weighed the appropriate options.

Secondly, your goal is to understand how the various strategy parameters play a role in the algorithm. As a rule, the more parameters are defined in a strategy, the more likely that overfitting is a problem. This isn’t always the case, and of course there is no magical number of parameters that is “just right”, but the key is to critically assess what fundamental value each one carries and how it contributes to the overall algorithm. Throughout the process, attempt to exclude parameters, and you may discover a simpler version of the algorithm. If that simpler version is able to accomplish the same goals, then you will have improved your chances of avoiding an overfitted solution.

One approach to parameter inquiry would be to ask the manager how you might expect the back-test results to change based on increases/decreases in each parameter’s proposed value. Once those sensitivities are tested, discuss how the expectations match up with the results. If any discrepancies occur, consider how the algorithm could function without that feature. Going back to the idea that the “best” back-tests are rarely any good, keep in mind that in theory given any data sample, there exists a set of many parameters and features that will allow a back-test to extract the maximum possible return. In other words, one can always add more conditions to an algorithm, thereby creating exactly the right decisions that would maximize gain in the observed period. But in practice, those gains are simply the effect of overfitting. Conversely, when removing overfitted features or parameters from a back-test, it’s reasonable to expect a decrease in the back-tested returns.

Sample Questions on Calculation Soundness

- Does the selected data sample reasonably represent the characteristics of the data that you can expect to materialize in the future?
- Was all available historical data included or were certain problematic periods left out?
- What frequency of data (daily/weekly/monthly observations) was used?

- Is the investment manager willing to provide detailed results of the same granularity or only summary results over longer observation periods?
- Are there claims being made regarding out-of-sample testing, and what type of claims?
- How many parameters are involved? Do they each serve a logical purpose?
- Does the manager have a solid understanding of the interactions and co-dependencies of various parameters?
- How would the parameters be altered if your investing goals were different?
- Can the strategy be simplified, but still meet your goals?
- Is the timing of available data, such as market prices, correctly reflected?
- Are trading costs and commissions, management fees, expenses etc. correctly reflected?

Can the Strategy Be Prudently Executed?

Finally, after your esteemed manager has successfully defended their back-testing practices, and provided a great deal of analysis and cooperation through the first two stages, only one question remains: can the manager actually implement the strategy? Your focus here should be on trading considerations, but also the manager’s procedures for model validation once live.

In order to assess what trading will be necessary to implement the strategy, ask the manager to provide an accounting of trades that would have been executed in the back-test versus daily traded volume for each security. If the strategy relies on even more timely trading, then a more granular approach is necessary. Aim for low participation in trading volume and inquire about the manager’s plans for engaging market makers to ensure that execution is cost effective.

Secondarily, if the strategy’s trading patterns are at all predictable, it’s important to ask what it would take for someone to willfully front run the strategy. “Front running” commonly refers to the practice of entering a trade prior to a different market participant, with the purpose of moving the price against them, and making a small profit. It’s obvious to see that if a strategy traded in a predictable, repeatable pattern, coupled with the need to trade a substantial volume of a particular security, a malicious person could take advantage of this and create a substantial and persistent cost to the strategy.

Sample Questions on Prudent Implementation

- What trading will be necessary to implement the strategy? How does it compare to historical traded volume?
- Could trades be predicted by a malicious outsider and subjected to front running?
- What is the manager’s experience in trading the relevant products/securities?
- Are there alternatives to these products and what situations would warrant a substitution?
- How will the manager validate model results?
- Are the manager’s investment operations organized in a diligent way that is likely to lead to quick discovery, escalation, and resolution of any mistakes or errors?

Conclusions

Unfortunately, there are not many shortcuts in the process of evaluating a complex quantitative investment proposal. Fortunately though, in the process of following such a lengthy and difficult examination, you are likely to gain an open and trusting partnership with the investment manager. Remember that when it comes to back-tests, the best are rarely any good, so the goal is a strategy that relies on solid principles and avoids the pitfalls of overfitting and hindsight. For that purpose, your main action items are:

- Make a plan for how you are going to monitor and assess the strategy results; and stick to it! When assessing the suitability of a back-tested strategy, resist the urge to shift into hindsight traps, and be diligent in documenting your investment rationale. Even though it may be tedious, try partitioning the back-test into small windows and assessing performance as you would with a live strategy. Focus on causes for over- and underperformance, and make sure you can stick with it.

- Diligently examine the data and the methodology used to construct the back-test. If historical data was omitted from the back-test, make sure there are good reasons for doing so. Question each parameter’s function and ask for sensitivity analysis to verify the correct understanding of that function. Throughout the process, attempt to reduce complexity by excluding features or parameters that have limited use.
- Ensure that the manager has a solid execution plan in place. Projected trades should be small compared to each security’s trading volume, and trade patterns and timing should not be easily deduced by outsiders. Finally, consider the investment managers procedures for review and validation of the models, as well as error escalation and resolution practices.