What Works on Wall Street

The Power of Back Testing Investment Strategies

Posted on September 25, 2014

Why should investors' back test an investment strategy? Investment advice bombards us from many directions with little to support it but anecdote. Many times, a manager will give a handful of stocks as examples, demonstrating how well they went on to perform. Unfortunately, these managers conveniently ignore the many *other stocks* that also possessed the preferred characteristics but *failed*. A common error identified in behavioral research on the stock market is this tendency to generalize from the particular, with evidence showing that people often "delete" from their memory those instances where they did poorly. This leaves them with the strongest memories centered on the few stocks that performed very well for them, and the faintest memory for those that performed badly. We therefore must look at how well overall *strategies,* not individual stocks, perform. There's often a chasm of difference between what we *think* might work and what really *does* work. We can also get information that conventional manager's lack: How often does the strategy beat its benchmark?

Jim O'Shaughnessy
Yahoo Finance Contributor

Jim O'Shaughnessy is the Chairman and CEO of O'Shaughnessy Asset Management (OSAM). Jim is the author of four books on investing, including What Works on Wall Street, a BusinessWeek and New York Times Business bestseller.
More ›

Search

Recent Posts

Getting the Most Out of Your Equity Investments

Why Selling a Big Position of Puts the Day Before the Crash of '87 was a great trade

Successful Active Stock Investing is Hard: Here are Seven Traits that I Believe are Required for Active Investors to Win in the Long Term

Some Great Books to Read on Vacation, Part 2: J Through Z

and how quickly did it bounce back? Knowing these facts helps investors remain confident of the strategy, particularly when it is underperforming.

In conducting back tests, mygoal is to bring a more methodical, scientific method to stock market decisions and portfolio construction. To do this, I have tried to stay true to those scientific rules that distinguish a method from a less rigorous model. Among these rules:

1) **An Explicit Method.** All models must use explicitly stated rules. There must be no ambiguity in the statement of the rule to be tested. There is no allowance for a private or unique interpretation of the rule.

2) **A Public Rule.** The rule must be stated explicitly and publicly so anyone with the time, money, data, equipment and inclination can reproduce the results. The rule must make sense and must not be derived from the data.

3) **A Reliable Method.** Someone using the same rules and the same database must get the same results. Also, the results must be consistent over time. Long-term results cannot owe all their benefit to a few years.

4) **An Objective Rule.** I have attempted to use only rules that are intuitive, logical and appeal to sensibility, but in all cases the rules are objective. They are independent of the social position, financial status and cultural background of the investigator and do not require superior insight, information or interpretation.

5) **A Reliable Database.** There are many problems with back testing, and the quality of data is the top concern. *All* large collections of historical data contain many errors. While Standard & Poor's Compustat

gold standard datasets for back testing, we must remain mindful of the limits of each. Undoubtedly, the databases contain stocks where a split was unaccounted for, where a bad book value persisted for several years, where earnings were misstated and went uncorrected, where a price was inverted from 31 to 13, etc. These problems will be present for *any* test of stock market methods and must not be discounted, especially when a method shows just a slight advantage over the market in general. For this edition we also use the CRSP dataset for the first time, which covers securities back to 1926.

Remember that the limits of the datasets are not trivial, and should be kept in mind as you review the results presented in this book. Edward F. McQuarrie published an article entitled "The Myth of 1926: How Much Do We Know About Long-Term Returns on U.S. Stocks?" in the Winter 2009 edition of *The Journal of Investing* in which he outlines some of the things to keep in mind when reviewing backtest results for various strategies. He points out that even comprehensive datasets like CRSP are faced with problems that include:

•Timeframe limitations: While the CRSP starts in 1926, McQuarrie notes that this still does not cover more than "50 percent of the historical record of widespread, large-scale stock trading in the United States, which goes back almost 200 years." Obviously, the monthly data from Compustat, starting in 1963, is even more limited in scope;

• Lack of coverage for all traded stocks: McQuarrie notes that "for more than 50 percent of its timeframe, the CRSP dataset excludes the majority of stocks trading in the United States, especially the smaller and more vulnerable enterprises."

upward bias to results, since many of the stocks added were added *because* they had been successful.

Thus, even though these datasets are among the best for analyzing the results to various styles of investing, it is important to keep their limitations in mind and contrast the results to those derived from other data series such as the Dimson, Marsh, Staunton Global Return Series featured in the book *Triumph of the Optimists: 101 Years of Global Investment Returns*, markets outside the United States such as those covered by MSCI and finally additional U.S. datasets such as the Value Line and Worldscope databases.

Potential Pitfalls

Many studies of Wall Street's favorite investment methods have been seriously flawed. Among their problems:

Data-Mining. It takes approximately 40 minutes for an express train to go from Greenwich, Connecticut to Grand Central Station in Manhattan. In that time, you could look around your car and find all sorts of statistically relevant characteristics about your fellow passengers. Perhaps there are a huge number of blondes, or 75 percent have blue eyes, or the majority was born in May. These relationships, however, are most likely the result of chance occurrences and probably wouldn't be true for the car in front of or behind you. When you went looking for these relationships, you went data-mining. You've found a statistical relationship that fits *one set of data very well, but will not translate to another*.As statisticians have been known to quip,if you torture the data long enough, it will confess to anything! Thus, if there is no sound theoretical, economic or intuitive, common sense reason for the relationship, it's most likely a chance occurrence. If you see strategies that require you buy stocks only on a confirm that the excess returns are genuine is to test them on different periods or sub-periods or in different markets, such as those of European countries. Indeed, we can look at the new results from the CRSP data between 1926 and 1963 as a validation of our previous findings. Research we have conducted in EAFE, which is the Europe and Far East Asia dataset maintained by MSCI, show the strategies performing with a similar level of excess returns as those in the United States.

Another technique that we employ is bootstrapping the data. Bootstrapping randomly resamples the overall results for the various strategies we test obtained by running 100 randomly selected subperiods to make certain that none of the randomly selected periods vary to any significant degree from the overall results shown for the various strategies. Typically, we view a factor as useful or predictive when there is a large spread between the annualized returns of the best and worst decile of that factor. The fact that the best decile of stocks with the best (highest) six-month price momentum beats the worst decile (stocks with the worst price momentum) by 9.96 percent per year for the last 83 years is powerful information that greatly influences how we advocate managing money. To eliminate any potential sample bias in this analysis we run a test on randomly selected sub-samples of the data to make sure that similar decile return spreads exist regardless of the group of stocks that we are considering. For each of the 100 iterations of each bootstrap test, we first randomly select 50 percent of the possible monthly dates in our backtest and discard the other 50 percent. We then randomly select 50 percent of the stocks available on each of those dates and discard the rest. This gives us just 25 percent of our original universe on which to run our decile analysis. We do this 100 times for each factor and analyze the decile worst decile remain consistent in these 100 iterations. Said another way, for the six-month price appreciation factor no matter which group of stocks are possible investments, it is always better to buy the decile with the best price momentum. If we discovered that there were large inconsistencies in the bootstrapped data, we would have less confidence in the results and investigate if there was any evidence of unintentional data mining inherent in the test.

A Limited Time Period. *Anything* can look good for five or even ten years. There are innumerable strategies that look great during some time periods but perform horribly over the long-term. Even zany strategies can work in any given year. For example, a portfolio of stocks with ticker symbols that are vowels, A, E, I, O, U and Y beat the S&P 500 by more than 11 percent in 1996, but that doesn't make it a good strategy! It simply means that in 1996, chance led it to outperform the S&P 500. This is referred to in the literature as the small sample bias, whereby people look at a recent five year return and expect it hold true for *all* five year periods. The *more* time studied, the greater the chance a strategy will continue to work in the future. Statistically, you will always have greater confidence in results derived from large samples than in those derived from small ones.

Survivorship Bias, or Then It Was There, Now It's Thin Air. Many studies don't include stocks that fail, producing an upward bias to their results. Numerous companies disappear from the database because of bankruptcy, or more brightly, takeover. While most new studies include a research file made up of delisted stocks, many early ones did not.

information was available when it was not. For example, researchers often assumed you had annual earnings data in January; in reality it might not be available until March. This upwardly biases results.

Rules of the Game

I have attempted to correct these problems by using the following methodology:

1) **Universe.** For this edition of the book, we use two datasets—the Standard & Poor's Compustat Active and Research Database from 1963 through 2009 and the Center for Research in Security Price (CRSP) dataset from 1926 through 2009. The S&P Compustat Database currently covers nearly 13,000 securities in North America and keeps historical records of financial and statistical information for the vast majority of traded securities–since 1950 for annual data and since 1963 for quarterly data. The CRSP dataset provides US daily corporate actions, price, volume, return, and shares outstanding data for securities with primary listings on the NYSE, NASDAQ, Amex, and ARCA exchanges. Both Compustat's and CRSP's research file includes stocks originally listed in the dataset but removed due to merger, bankruptcy or other reason. This avoids *survivorship bias*. I cannot overstate the importance of testing strategies over long periods of time. Any study from the early 1970s to the early 1980s will find strong results for value investing, just as any study from the 1960s and 1990s will favor growth stocks. Styles come in and out of fashion on Wall Street, so the longer the time period studied, the more illuminating the results. From a statistical viewpoint, the strangest results come from the smallest samples. Large samples always provide better conclusions than small ones. Some pension consultants use a branch of statistics called *reliability mathematics* that use past returns to predict future performance. They've found

2)**MarketCapitalization.** Except for specific small capitalization tests, I review stocks from two distinct groups. The first includes only stocks with market capitalizations in excess of $200 million (adjusted for inflation), called "All Stocks" throughout the book. The second group includes larger, better-known stocks with market capitalizations greater than the database average (usually the top 17 percent of the database by market capitalization). These larger stocks are called "Large Stocks" throughout the book.

In all cases, I remove the smallest stocks in the database from consideration. For example, at the end of 2009, of the 6,705 stocks in our dataset, more than 2,555 stocks were jettisoned because their market capitalization fell below an inflation-adjusted minimum of $200 million. In the same year, only 651 stocks had market capitalizations exceeding the database average. We also remove stocks that appear in the Compustat but have no market capitalization, duplicate issues, shares of mutual funds, etc.

I use the $200 million minimum to avoid microcap stocks and focus only on those stocks that a professional investor could buy without running into liquidity problems. Inflation has taken its toll: A stock with a market capitalization of $29.40 million in 1963 is the equivalent of $200 million stock at the end of 2009. The same $200 million deflated back to 1926 was the equivalent of a $16.8 million stock.

Eliminating micro-cap stocks considerably reduces the returns for several of the factors we will study. It also puts the results featured in *What Works on Wall Street* at a disadvantage when compared with many academic studies that include microcap stocks. We have found that by eliminating micro-caps, our results appear to be significantly lower than those of results that are far more likely to be able to replicate in the real world. Microcap stocks possess virtually no trading liquidity, and a large order would send their prices skyrocketing. Thus, while it is easy to *assume* that you could purchase and sell these securities at their listed price in the historical dataset, I believe that is an illusion and unnecessarily gives an upward bias to the results of studies that allow their inclusion.

3) **Avoiding Look-Ahead Bias**. We use only publicly available monthly information. To ensure that we are not selecting stocks based on information that is *not* publicly known, we lag quarterly data by three months and annual data by six months. While this can have the effect of making information slightly stale, it is necessary to avoid look ahead bias.

A properly conducted back test can give investors insights that those who base their buying and selling of securities on stories or news accounts will never have—and it will give them important information that conventional investors lack. In short, it can give you a true edge and allow you to remove emotions and “arbitrage human nature.”

marketonomics liked this

nachopugliese liked this

myoryol liked this

andrewnyquist liked this

randyranks liked this

carsondahlberg liked this

financecontributors reblogged this from jimoshaughnessy

jimoshaughnessy posted this

PREV POST NEXT POST

Home O'Shaughnessy Asset Management About Archives Disclaimer