NORTHFIELD | modeling financial markets worldwide

Back-Testing: Useful Tool or "Financial Charlatanism"

Dan diBartolomeo

Northfield Annual Summer Seminar, Newport

June 2016

Name That Author?

Even trained statisticians often fail to appreciate the extent to which statistics are vitiated by the unrecorded assumptions of their interpreters... It is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life and confers comparative immunity from disease. A university degree, a daily bath, the owning of thirty pairs of trousers, a knowledge of Wagner's music, a pew in church, anything, in short, that implies more means and better nurture... can be statistically palmed off as a magic spell conferring all sorts of privileges...

The mathematician whose correlations would fill a Newton with admiration, may, in collecting and accepting data and drawing conclusions from them fall into quite crude errors by just such popular oversights as I have been describing.

Introduction

- In 1995, I did a lecture at a CFA event at Northwestern
  - The title was a question "How to Blow a Back-Test?"
  - Sadly, the answer was "Believe the Results"
- In the twenty years since that lecture, industry practices around "back-testing" have not changed much.
  - Such simulations remain a staple activity of investment firms.
  - Favorable back test results are an essential part of the marketing of new portfolio management strategies.
- Conventional back-testing procedures almost always yield results that are statistically insignificant and have little real content.
  - If anyone tells you to believe the results from a typical back test, they are *either a liar, a fool or both.*

Presentation Outline

- The first part of the presentation will deal with the conceptual limitations of typical back-testing and advocate for a "best case scenario" interpretation of the results.
- Next we will discuss the statistical properties of historical sample data and degree of assumption dependence associated with "the future will be similar to the past".
- We will then review three key papers showing the weaknesses in conventional tests.
- The remainder of the presentation will describe four ways to improve the validity of back-tests

Back-Testing

- Bailey, Borwein, de Prado, and Zhu (AMS, 2012) define a back-test as "a historical simulation of an algorithmic investment strategy".
  - The phrase "Financial Charlatanism" in their title.
- A back test is useful and valid when the *correctly formulated expectation* of out of sample performance is equal to the in-sample simulated performance. This almost never happens.
  - Stationarity assumptions and look ahead bias leading to invalid priors
  - Unrealistic test conditions (e.g. zero trading costs)
  - Insufficient sample size and overfitting
  - Incorrect statistical interpretation of outcomes

Common Sense Problems

- The whole concept of back-testing rests on the idea that history can be our guide to the future.
  - There are extreme assumptions that the passage of time in financial markets is a stationary process. We assume that to a meaningful degree, future financial outcomes will be like the past.
  - Kahn and Rudd (FAJ, 1995) show that this is a very weak assumption
- Almost all back-tests are look ahead biased either in raw data or strategic concepts
  - "If I knew then what I know now, we would have performed well"
  - The physicist Steven Hawking said that one of the proofs that time only goes forward is that "otherwise, we could invent a computer that would report tomorrow's stock prices"
  - Everybody watches everybody else to see what's working, much like a fashion show.

More Common Sense Problems

- Back-tests assume that a given financial market participant could have traded without any alteration in the course of past events.
  - Given that every buyer must have a seller and vice versa this idea is silly on it's face. It is doubly silly for large institutions.
- Most quantitative models for investment management rely on a series of parameters to formulate return expectations or do portfolio construction tasks.
  - With the widespread availability of cheap computer power, there is a overwhelming urge to carry out many tests so as to find the best combination of the parameters.
  - The greater the number of parameters and the extent of parameter "fitting", the greater the length of the sample period needed to validate the process. "Overfitting is rampant"

Even More Common Sense Problems

- Most back-tests seek to simulate a period of history based on success concepts to which our attention has been but that could only be known after the fact.
  - If we have a million monkeys pounding on a million keyboards, and one of them types out the first 10 pages of Hamlet, it does not mean that this (or any) monkey will be the next Shakespeare.
  - See Kahn (FAJ, 1997)
- Conventional tests have extremely low statistical power
  - Most tests simulate over the history that was actually experienced. There is rarely any testing of how a strategy would have done over the infinite number of other paths the evolution of history might have taken.

Defining the Parameter Space

- Bouchard, Laloux, Cizeau and Potters (MMMAS, 2000)
- When building and testing a quantitative model of investment returns it would be really good to know how many parameters you are looking for to explain what you have observed.
  - They create long time series of simulated security returns using a random number generator
  - They then form the covariance matrix of these random series and do eigenvector decomposition (principle components analysis) to see how many spurious eigenvectors they get
  - Depending on the number of simulated securities and the number of periods of simulated returns they often get many, when we know in advance the correct answer is always zero.

"Deflategate: A Random Back-Test?"

- A paper by Novy-Marx (2014) extends the use of randomly generated "signals" to pick stocks
  - Generate random signals and correlate each set of random signals to stock performance.
  - Select several of the signals with the best fit in sample
  - Combine the selected signals into an "alpha" signal and scale to the expected distribution of returns
  - You get massively efficient fit to in-sample data while the correct expectation is zero predictive power out of sample.
  - Since we don't think our real world alpha signals are totally random, the paper provides a way to analyze the extent to which out of sample results should be "deflated" to form realistic expectations of predictive power.

Mathematical Analysis of Overfitting

- Bailey, Borwein, de Prado, and Zhu (AMS, 2012) do a very detailed analysis of the mathematics of "overfitting"
  - Most investment models have multiple parameters. The greater the number of parameters, the larger the in-sample period must be to validate the process.
  - A formula is provided showing that you need really long tests for the expectation OOS performance to converge to in sample (IS)
  - Still assumes stationarity and distributional sufficiency.
- You can't get around this problem by simply reducing observation periods from months to weeks to days to get more observations
  - Basic assumptions of IID processes break down completely at the shorter end of the time scale.

Do We Need Back-Testing at All?

- Sneddon (Northfield Conference 2005) and (JOI, 2008) provides a way to analytically forecast the long term performance (alpha, Sharpe ratio, etc.) of a strategy conditional on the forecasting power (IC) of the "signal".
  - Given whatever you believe your IC will be, the long term performance is predictable given certain assumptions.
  - You have to assume your IC is positive but can vary over time, and that alpha signals decay with time so your IC at a one month horizon may be different than at a six month horizon.
  - Market impact portion of transaction costs is linear in trade size
  - The security covariance matrix for risk is accurate
  - We form optimal active portfolios that are sufficiently diverse that transfer coefficients are high or at least predictable

If We Have to Back-Test, Let's Do it Better

- Look ahead bias can be substantially reduced by using databases that provide "point in time data" so you are seeing information (e.g. financial statements) as they would have been at the time.
  - See Bogue (Northfield conferences, 1995 and 2001)
- Don't just look at simple simulated performance
  - Use a fully detailed attribution system to find results that are driven by specific events (e.g. the tech bubble) a subset of securities within the history. Performance driven by narrow causes is even less likely to be stable.
  - Check the statistical significance of all aspects of the back-test, suitably deflating T stats and P values for the number of alternative parameter configurations. See Harvey and Liu (SSRN, 2013)

Change the Basic Metric to the EIC

- diBartolomeo (JPrfMs, 2008) provides an alternative measure for investment performance, the "Effective Information Coefficient" (EIC)
  - The effectiveness of an investment strategy is measured at the individual position level, not the portfolio return level, massively increasing sample size.
  - EIC requires that every portfolio manager must believe that their portfolio is optimal or they would do something else.
  - Every security position must contribute enough alpha to offset it's marginal contribution to risk (definition of optimality)
  - If you assume that the security covariance matrix for risk is well defined you can generate implied alphas which can be correlated to subsequent outcomes (EIC). Automatically incorporates constraints and trading costs.

Testing For the Fragility of Back-test Results

- Conventional back-test results only simulated what would have happened with the history we actually experienced. We want to understand how outcomes would have changed if history had been different.
  - A statistical procedure called bootstrapping can be used to randomize a sequence of historical events.
  - We can simulate thousands of paths of possible historical sequences in random order, with some periods skipped, or one time events occurring more than once.
  - History can be filtered (e.g. only study periods when the US economy was in recession)
  - "What if" scenarios can be overlaid (e.g. adjust event probabilities so interest rates drop 3% over the life of every path)

Conclusions

- Back-tests as usually done in the investment industry are grossly insufficient as indications of probable future out of sample performance.
  - At best, conventional tests should be indications of the likely upper bound of out of sample performance, not the expected value
  - There is an expanding statistical literature showing how extensive a conventional test must be to have a reasonable degree of legitimacy, and how to deflate the built in optimism.
- There are better ways to go
  - The multi-period formulation of portfolio dynamics by Sneddon provides a convenient and powerful alternative to back-testing.
  - A combination of detailed performance attribution, EIC analysis, and bootstrapping can provide a more informative and robust analysis. Obviously all these tools are available from Northfield.