Effective Testing on Portfolio123 Testing (backtest, rank performance test, simulation)

Marc Gerstein - 4/15/16

A Common But Potentially Dangerous Use

- Empirical Study – testing to determine, from scratch, whether an idea is sound
  - If I create this ratio, (x+y)/z and use it as a basis to rank stocks, will it work?
    - If so, I next ask whether my test "robust" according to statistical "best practices"
    - If yes – Great! I can and will use it in a model.

The Challenge in Financial Testing

- We don't "interpolate"
  - Interpolation involves modeling relationships observed within in a population (or within a properly constructed sample), in order to put those relationships to use *within the same population*
- We "extrapolate"
  - This involves modeling relationships observed within a population or sample in order to put those relationships to use *in a completely different population*

Example: Interpolate vs. Extrapolate

- Interpolation – Medical research
  - If such-as-such chemical compound is ingested by humans having such-and-such disease, the disease will disappear and without adverse side effects.
  - If a relationship is demonstrated, and if it is robust (which is where companies fight it out with the FDA), then we know the compound can be safely and effectively administered among humans
- Extrapolation – A Whole Different Ballgame
  - That study would not support use of the same compound to cure dogs
  - Dogs and humans are similar in some ways but different in other ways (example: chocolate is a popular treat among humans, but it can kill dogs)

Approaching Extrapolation

- One way to extrapolate is to do a different study on each possible population
  - E.g., study the same chemical compound again on dogs
- Another approach is to focus the initial study on characteristics of the first population that we can reasonably assume will carry over into other populations

The Two Populations on Portfolio123

- Population 1 – The past, the point in time database ("in sample")
  - We can with 100% certainty identify and model relationships that are valid within this population
  - Everybody does it
- Population 2- The future, live money (out of sample)
  - Because we don't (can't) have data from the future, we're forced to use the second approach to extrapolation – working with concepts we can reasonably expect to carry over from Population 1 (the past) to Population 2 (the future)

Practical Reality

- What happens if we aim our testing entirely at interpolation?
- It might work
  - This will happen if the future resembles the past
  - This can happen for a short time, it can happen over the course of many years, or it can come and go at what the user perceives to be random intervals or based on hot-and-cold fads
- It might not work
  - This will happen if the future differs from the past in a manner that is relevant to the model
    - The better the interpolation – the more tightly tuned the model is to the past – the more opportunities there are for the future to differ
    - This is why 5-stock 90% Alpha sims are more likely to disappoint than 20-stock 15% alpha models
    - Robustness is irrelevant; all robustness means is that you did a superior job interpolating
  - The user will experience this as disappointing out-of-sample performance

How to model for Extrapolation

- Test ideas you Reasonably expect to be capable of operating in both the past and the future
- William O'Shaughnessy, in *What Works on Wall Street*, echoes this with a warning that actually provides a hint as to how we might accomplish it:
  - "If there is no sound theoretical, economic, or intuitive, common sense reason for the relationship, it's most likely a chance occurrence."
    - Tortoriello follows this same principal and works the same way
    - The key isn't in what they test – it's what they don't bother to test (things they know ahead of time lack the economic, intuitive or common sense basis)!

The Solution

- Work with and test ideas you reasonably expect to impact stock prices
  - Don't focus on trends or what's happening now or what happened in the past
  - Focus on ideas you know, ahead of time and without testing, should work
- This requires that there be a foundational idea about how stocks are priced and why they move
  - Good news: Such a foundation exists and is easy to understand
  - Bad news: The idea is brutally difficult to apply which is why we don't all become big winners on every trade
  - Testing is used to provide feedback on the efficacy of our implementation ideas

The Single-Biggest Error

- People who test improperly, who interpolate, argue that stocks behave irrationally and that there is no such thing as “a foundational idea about how stocks are priced and why they move”
- They are wrong, very badly wrong
- Such an idea does exist and is well known
- The key for proper testing, testing that could be used to extrapolate (which is what we must do) is understanding the foundational idea, how we can apply it to the real world and how we can use the adaptations as the basis for our models

The Foundation: the Dividend Discount Model (DDM)

- It tells us that the fair price of a stock is equal to the present value of all the money we expect to receive as a result of owning it:
  - Dividends
  - Proceeds from an eventual sale
- THIS ABSOLUTELY POSITIVELY MUST BE TRUE!
  - To argue against it would mean advocating in favor of something like an immediate exchange of $100 for $5; nobody would do that
- Implementation Problem: How do we know when we'll sell and what we'll get
  - Answer: Mathematicians restate the idea assuming we'll hold to infinity, thus allowing us to focus on dividends only

The DDM Formula

- P = D / (R – G)
  - P = fair price
  - D = dividend
  - R = required rate of return
  - G = expected (through infinity) dividend growth rate

  – We can also use a formula for R (the Capital Asset Pricing Model or "CAPM")
    - R = RF + (B * (RM – RF))
      – RF = Risk Free (treasury) rate
      – RM = expected return of the equity market as a whole
      – B = Beta, the indicator of stock's volatility relative to the market

Practical Challenges

- We can't literally use DDM
  - Many stocks don't pay dividends
    - The idea here would be to do a DDM valuation as of some forecasted future date when dividends are expected to be introduced and then apply the answer to today using a "present value" computation
  - The main problem is that it's too difficult generate credible inputs
    - The worst problem is with G, must be an infinite growth rate that presumes a very mature company (i.e. we have to make sure G is less than R lest we wind up with a negative fair price)

Coping Mechanism

- "It is better to be vaguely right than precisely wrong."
  - Craveth Read (British logician)
- That works for us since we're dealing with the future and, thus, have no hope of being precisely right . . . So we might as well aim for vaguely right (or as Prof. Asawath Damodaran says, less wrong then everybody else)
  - Data mined "interpolated" sims, on the other hand, unwittingly aim to be and usually wind up being "precisely wrong."

Important Approximation # 1

- Since we know we cannot precisely implement DDM, we do the next best thing
- We adapt, adjust, approximate, etc. using ideas we reasonably believe will push us closer in the direction of fair DDM valuation
  - For example, we consider value because a stock that it more closely aligned with current earnings is more likely to be reasonably aligned with the stream of future dividends, since dividends come from earnings

Important Approximation #2

- We don't literally have to restrict ourselves to the relevant set of terms: D, G, R (including RF, RM and B)
- We can, instead, substitute other ideas that are plausibly related to these terms.
  - For example, we can consider financial strength since the balance sheet impacts the stability of profitability which impacts the stability of the stock (B) which in turn impacts R

Making Sense of the Approximations

- The approximations, packaged together, make up the body of "fundamental analysis"
- The fundamental ratios etc. that are widely used did not become so just for the heck of it or because so-and-so said so
- Fundamental ratios get to be widely used because of the way they logically connect us to the goal; alignment with the ideal (albeit, sadly, incalculable) DDM ideal
- "Empirical" data miners might hit, through luck, on factors or formulas that meet such criteria. But often their 20-20 hindsight leads them to rely on formulas that do not qualify on this basis
  - Use of factors that just happened to have worked without regard to whether they SHOULD work (as per these criteria) is what causes the super sims to fall apart when applied to the future

Example: Value and PE

- Substitute E (EPS) for D
  - We can get away with this because we see that the market behaves consistently with an expectation that all company earnings accrue directly to shareholder and that the latter implicitly choose to reinvest all or a lot of it back into the company
- So now, P = D/(R-G) becomes P=E/(R-G)
- Algebraic reshuffling: P/E = 1/(R-G)

We've Unlocked the Key to Value Modeling

- P/E = 1/(R-G)
  - Because G is a negative number in the denominator, we know that as G goes up, so, too does P/E
    - The inspiration for the PEG ratio! But PEG isn't the whole story . . .
  - Because R is a positive number in the denominator, we know that as R goes up, P/E goes down
    - Interest rates (RF) is a major component of R; that's why falling interest rates are good for stocks (the push P/E up) and vice versa
    - B (beta) is also a major component of R, the only company-specific component; that's why less risky stocks are often criticized as being expensive and why deep values can be high risk – that's how they are supposed to be priced

It's not just P/E

- Price/Sales
  - P/S = Margin / (R-G)
    - Stronger margins and better sales growth push valuations up
    - S relates to E and E relates to D
- Price/Cash Flow
  - P/CF (or P/FCF) = 1/(R-G)
  - Assumes a similar relationship between CF, FCF, etc. and D as we assumed with E
- Price/Book
  - P/B = ROE/ (R-G)
  - This is because E = ROE * B

This is What Fundamentals are all about

- Extensions of DDM in recent slides have already opened the door to quite a few fundamentals
- We keep going
  - Turnover is a component of ROE as is margin
  - Debt, liquidity ratios, etc. relate to risk and hence B and therefore R
  - ROE, ROI, and ROA mare indicators of persistent profit growth capability and hence relate to B and R (and of course G)

It's more than just fundamentals

- Because we're looking into the future, we have to be creative in our quest for clues
  - Sentiment-related and technical factor or formula becomes valuable if it passes this test: the number is the way it is because "they" are assuming . . . .
    - "They" could be analysts and/or the market as a whole
    - This is an important source of information because when we look into the future, we need to be open to qualitative judgments that cannot be expressed in terms of historical fundamentals

It's also a lot more than "fair price"

- Robert Schiller and Charles Lee
  - P != V
    - (We can't assume Price is equal to Value)
  - P = V + N
    - Price is equal to Value plus Noise
- N can be understood in ways we can model
  - N moves based on sentiment (fads, trends), and also based on information availability or lack thereof that makes it hard or easy to credibly value a stock
  - Wal Mart is pretty clear-cut; it can be valued, so N is small
  - Biotech microcaps are brutally difficult to value so their prices are all N all the time
- Sentiment and Technical analysis help us get a handle on the ebbs and flows of N

Making it Work

- There's a lot we can use to help us identify potentially good stocks
- But . . .
  - We need to remember we're dealing with the future, so we have to come up with relationships we think will be sustainable
  - We also need to express our ideas in ways a computer can read and process
    - We can't simply say we want companies with good growth prospects
    - We have to specifically define "good growth prospects" in p123 language – which can incorporate historical fundamentals, sentiment, and/or technicals

Building and Testing a Value Model

- We can start with a Value Ranking system that includes one or more ratios – sorted assuming lower is better
- We could try stopping there and it may work in test, but we know we're missing things
- So we'll add factors relating to G (higher is better) and others that influence future B (lower is better)

Model Design

- What is in the ranking system and what is in the screen?
  - Assume we use a ranking system for Value
  - Assume we use a screen (buy rules) to prequalify the to-be-ranked universe on the basis of companies with good growth prospects and/or high quality (we can and do assume that higher quality companies are likely to have more stable earnings and, hence, lesser future Betas)
- However we configure the model, we look for stocks with ratios that are lower than we think they would be if the market is correctly understanding G and R
  - What we're really doing here is information arbitrage

What We Test - 1

- Have we done a good job specifying G
  - Sales5YCGr% ?
  - Sales3YCGr% ?
  - (Sales%ChgTTM-Sales5YCGr%)/abs(Sales5YCGr%) ?
  - EPS items?
  - LTGrthMean ?
  - (NextFYEPSMean- NextFYEPS4WkAgo)/ NextFYEPS4WkAgo ?
  - AvgRec?
  - Etc., etc., etc.

What we Test - 2

- Have we done a good job specifying Value?
  - PEExclXorTTM ?
    - Are we getting too many bad numbers due to inclusion of special items?
  - Pr2CashFlTTM ?
    - It doesn't allow for the capital spending equivalent of depreciation: Is this a problem?
  - Pr2BookQ ?
    - With so many unquantifiable assets nowadays, does this ratio still work?
  - Etc., etc., etc.

What we Test - 3

- How should we specify Risk/Quality?
  - Does ROE do it, or are we better off with ROI or ROA?
  - Is 5Y good, or TTM?
  - Should we compare them to get an indication of ROE trend?
    - Can ROE serve as a proxy for growth – there is a relationship
  - Are debt ratios useful, or perhaps interest coverage and should these be industry comparisons?
  - Earnings quality is an indicator of persistence and risk

What we Test - 4

- Are we being impacted by "specification error?"
  - We expect that a stock with a PE of 12 is better valued than a stock with identical G and R characteristics but a PE of 20
  - But maybe not!
    - What if the 12 PE is calculated on the basis of EPS that is temporarily inflated by a one-time never to be seen again gain and that an EPS more reflective of the company's underlying fundamentals would lead to a PE of 35
    - This sort of thing happens all the time
      - Don't expect Xor to bail you out; accounting standards setters have pretty much defined Xor out of existence
    - So for us, and contrary to general statistical best practices, more factors can be better than fewer, as we screen out high probability sources of specification error or diversify them away (with, say, five value ratios instead of one)
  - It's important to hunt for specification error by running your model and sampling passing companies, to see if they satisfy the spirit of the law
  - Specification error is a huge source of disappointing live-money results
    - It's impractical to expect to completely eliminate it, but we can and should try to do the best we can to reduce it

What we Test - Summary

- There are countless things we can test but everything is related to a common goal:
  - Have we effectively specified our ideas?
    - Have we succeeded in translating our ideas into P123 language?
    - Have we captured relationships likely to persist into the future?
    - Have we dealt with specification errors as best we can?
- Because the questions we ask are so open-ended, it's very risky to implement a model without testing to see if our specification is plausible
- It's important to understand we are not on an empirical treasure hunt for what works.
  - The choice of what to test or what to discard even without test is critical.
  - We can only test ideas we have reasons to expect will work, and look at the test to assess the efficacy of the specification
  - Data miners go bad testing things they shouldn't test, the kinds of things O'Shaughnessy and Tortoriello would reject without even bothering to test

What we can infer from Testing

- A successful test is one that allows us to assume that the relationships we specified have a good probability of identifying stocks with the potential to outperform the crowd (however we define – benchmark – it)
  - That's what we should have been looking at in our test results; excess return, alpha, etc.

What we Cannot Assume from a Test

- We can never assume that the stock will go to a particular price
  - Market action is the single biggest component of any stock price, so much of what a stock will do depends on the market
- Because P123 doesn't have tools to predict future market prices, we cannot predict target prices
  - It's not clear anybody has anything that helps with this
  - This is an important reason why analysts don't take "Target Prices" seriously and why many sites don't provide them
    - They might, however, be used to construct a sentiment indicator