Backtesting:
A Practitioner’s Guide to Assessing Strategies and Avoiding Pitfalls

March 5, 2015
2:15 pm – 3:30 pm

Olivier Sarfati, Citigroup

Many thanks to Joseph Kogan and Alexander Shapiro for helping with this presentation
Favorite Quotes about Backtesting

“I only believe in the backtests that I doctored myself.”
- Paraphrased from Winston Churchill

“Lies, damned lies…and backtests.”
- Paraphrased from Benjamin Disraeli

“I’ve never seen a bad backtest.”
- Unnamed Equity Derivatives Professional
Introduction
Introduction

• Too many investment strategies do not live up to expectations: they perform well in-sample, but fail to do so when implemented.

• Having dedicated our professional lives to investments, we are convinced that opportunities exist as a result of asymmetric information, risk preferences, tax treatments, barriers to entry and other structural imbalances.

• A good backtest should guide investors to those opportunities, and help them decide which strategies to privilege from an asset allocation standpoint.

• This is an attempt at presenting some of the mathematical concepts to avoid common pitfalls
What is a Backtest?
Summary:

Practitioners and academics test many strategies over the same historical dataset; however, their statistical tests of significance and their estimates of expected future Sharpe ratios do not account for this over-testing, thereby returning misleadingly positive results. As a consequence, “most of the empirical research in finance...is likely false.”

Solutions include:

• Using more stringent tests and thresholds for significance, such as Family-Wise Error Rate testing, and False Discovery Rate testing.

• Adjusting expected Sharpe Ratios downwards to account for multiple testing.
The scientific journal “Basic and Applied Social Psychology” has banned the use of significance testing because:

• It is widely misused – practitioners assume that a p-value is the probability of a null hypothesis given the data. In reality, it is the probability of the data given the null hypothesis.

• It is overused, often seen as the main test of whether a scientific claim is true. However, it is easy to achieve a spuriously good result on this test, leading to skewed incentives for researchers to datamine, i.e. “torture the data until it confesses”.

Source: Editorial David Trafimow, Michael Marks Basic and Applied Social Psychology Vol. 37, Iss. 1, 2015
What are Backtests and Why do People Run them?

**Backtest** (bæk-tɛst) *n.* a simulation that seeks to estimate the future performance of an investment strategy by testing its performance in the past

- Assess a complicated strategy by putting a value on it
- Underlying principle: “If it worked and things haven’t changed too much it’ll keep working. And If it didn’t work before, why bother?”

What are Backtests and Why do People Run them?

**backtest** (bæk-tɛst) *n.* a simulation that seeks to estimate the future performance of an investment strategy by testing its performance in the past

- Assess a complicated strategy by putting a value on it
- Underlying principle: “If it worked and things haven’t changed too much it’ll keep working. And If it didn’t work before, why bother?”
- **But then why do backtests always include the disclaimer that “past performance doesn’t guarantee future returns”**?

Past Performance Doesn’t Guarantee Future Returns

A Top-Selling Fixed Income Annuity...

• ALTVI is a competitor’s alpha strategy
• Comprised of 24 different future contracts, commodities, rates, and currencies
• Issued in 2012

**Backtest (in sample)**

The ALTVI is an excess return index, thus we compare it to the excess return on the S&P 500.
Past Performance Doesn’t Guarantee Future Returns

A Top-Selling Fixed Income Annuity...

- ALTVI is a competitor’s alpha strategy
- Comprised of 24 different future contracts, commodities, rates, and currencies
- Issued in 2012

**Backtest (in sample)**

Up 7.2% per year in up and in down markets

**Live (out of sample)**

Down 20% since peak

The ALTVI is an excess return index, thus we compare it to the excess return on the S&P 500
10 Pitfalls of Backtesting
The Pitfalls

1. Focusing on Expected Value, Ignoring other Metrics
2. Underestimating Survivorship Bias
3. Confusing Correlation and Causation
4. Selecting an Unrepresentative Time Period
5. Accounting for Outliers Improperly
6. Overfitting and Data Mining
7. Overlooking Mark to Market Performance
8. Failing to Expect the Unexpected
9. Underestimating Hidden Exposures
10. Forgetting Practical Aspects

Montparnasse Derailment, 22 October 1895
Source: http://en.wikipedia.org/wiki/Gare_Montparnasse
Pitfall 1: Focusing on Expected Value, Ignoring other Metrics
Focusing on Expected Value, Ignoring other Metrics

Pitfall

- Even if accurate, expected value alone does not completely describe a set of data

Example 1

- If you short a var swap, you might be able to estimate your payoff...on average
- But, due to the convexity, your max loss is much larger than your max gain

Solution

- Especially when looking at convex payoffs, such as for var swap or options strategies, focus on volatility, skew, and max drawdown.
- Don’t rely solely on statistics which are more appropriate to linear payoffs (such as average return)
WSJ Claims, “September Cruelest Month” – Short Sep?

Example 2

• Chart suggests Sep is the only month to have consistent negative returns

• Should you short Sep given this result?

http://online.wsj.com/articles/some-stock-strategists-brace-for-september-swoon-abreast-of-the-market-1409592143
The WSJ chart implies high probability of loss in Sep

Example 2 (Cont’d)
- So, Sep has lowest average return of all months
- What is the risk?
- What does the distribution look like?

What WSJ implies the distribution is:
- \( \sigma = 0.5 \)
- \( \sim 88\% \)
- \(-0.6\% \)

What the distribution actually is:
- \( \sigma = 4.5 \)
- \( \sim 55\% \)
- \(-0.6\% \)
Looking at data the correct way makes the difference

Solution

• Compute std. dev. and Sharpe ratio
• Graph data to visualize the distribution

% Returns Since 1950

By the way:
S&P -1.6%
this Sep
Pitfall 2: Underestimating Survivorship Bias
Underestimating Survivorship Bias

Pitfall

- Excluding failed companies from a backtest simply because they no longer exist
- This can skew results higher
Changes in Index Membership

Example 1

- Sell puts systematically on S&P names
- Backtest is prone to survivorship bias: S&P is periodically rebalanced, so not all current S&P members were in the index 14 years back

Solution

- In a backtest based on index constituents, account for changes in the index’s membership over time

% Return Since 2000

Index-weighted SPX members as of Sep 2014

Regular SPX

~400%age pt difference in return
Mutual fund performance: the ones that didn’t make it still count

Example 2

- Simple strategy: invest your savings in mutual funds
- Assess the returns of currently-existing mutual funds over a historical period
- Analysis will omit funds that were liquidated or merged away due to poor performance

Solution

- Remove bias by including all mutual funds that have existed over time
- Estimate percent market share of failed funds annually; estimate final value before redemption; add performance of failed funds to previous analysis

Over half of the funds introduced since 1997 have been liquidated

Graph illustrates performance decrease when bias is removed

<table>
<thead>
<tr>
<th># of L/S &amp; Market-Neutral Funds Introduced per Year</th>
<th>Source: Wall Street Journal, 8/6/2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liquidated or merged</td>
<td>Still operating</td>
</tr>
<tr>
<td>1997 '98 '99 '00 '01 '02 '03 '04 '05 '06</td>
<td></td>
</tr>
</tbody>
</table>

Standard vs. Bias-Adjusted Performance

<table>
<thead>
<tr>
<th>Long/Short</th>
<th>Market-Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Adjusted</td>
</tr>
<tr>
<td>10 years</td>
<td>15 years</td>
</tr>
<tr>
<td>5.05%</td>
<td>1.95%</td>
</tr>
<tr>
<td>3.42%</td>
<td>0.59%</td>
</tr>
</tbody>
</table>
Pitfall 3: Misunderstanding Correlation and Causation
Are fund flows predictive of market performance?

**Pitfall**

- Mistaking the directionality of a causal relationship
- Omitting a confounding variable that actually drives a relationship
- Assuming correlation implies the existence of causation

**Example 1**

- Strategy: use fund flows as a predictor of future market performance
- Idea is simple: managers get inflows, they have to invest → market goes up
- True?

**Forward-Looking:**

3m Equity Fund Flows* vs. Next 3m Perf.

Source: Citi Research

[Graph showing scatter plot with correlation coefficient R² = 0.0002]
Fund flows are at best coincidental with market performance

Solution

• Check to see if the direction of causality is the other way around, or if the relationship is coincident.
• Data suggests the relationship is coincident at best, with $R^2 = 0.26$.

Coincident:

3m Equity Fund Flows* vs. Last 3m Perf.

Source: Citi Research, EPFR. *Based on global equity flows.

Source: http://xkcd.com/552
Correlation vs. Causation: Watch out for confounding variables

Example 2

• Spurious association:
  - Distressed Lending → Bankruptcy

• Actual association driven by a third, confounding variable:
  - Insolvency
  - Confounding variable
  - Distressed Lending
  - Bankruptcy

Solution

• Recognize, assess, and control for possible confounding variables
• Verify that the alleged cause temporally precedes the effect
• Conduct cross validation to test for biases that may be caused by the existence of a confounding variable
“Rock Music Quality is a 4-Yr Predictor of U.S. Oil Production”

Example 3

- Rock music quality is correlated with US oil production
- Did the Beatles boost oil production?

Solution

- Devise a theory that accounts for the causal relationship by describing its various implications
- Test the theory with common sense

Source: Rolling Stone Magazine, US Department of Energy
Pitfall 4: Selecting an Unrepresentative Time Period
Selecting an Unrepresentative Time Period

Pitfall

- Time period is not just too long or too short, but
  1. is unrepresentative of the current environment
  2. lacks relevant market regimes
  3. does not include enough data points
  4. captures extreme historical events that are unlikely to repeat
The time period used can make all the difference...

**Example 1**
- Short VXX: backtest shows excellent results!

Source: Bloomberg
The time period used can make all the difference...

Example 1

- Short VXX: backtest shows excellent results!
- ...since 2012
- Time period is too short
- Looking back further shows effect of another regime: the crisis

Source: Bloomberg
The time period used can make all the difference...

**Example 1**
- Short VXX: backtest shows excellent results!
- ...since 2012
- Time period is too short
- Looking back further shows effect of another regime: the crisis

**Example 2**
- Play mean reversion on Brent vs. WTI – it’s a low vol spread

---

Source: Bloomberg
The time period used can make all the difference...

**Example 1**
- Short VXX: backtest shows excellent results!
- ...since 2012
- Time period is too short
- Looking back further shows effect of another regime: the crisis

**Example 2**
- Play mean reversion on Brent vs. WTI – it’s a low vol spread
- until 2011
...so choose your time period carefully

Solution

• Include a crisis, a low vol period, a period comparable to now; the last 20 years saw all three
• Isolate and understand results for every regime
• Search for one-off historical explanatory factors
• If strategy underperforms in a certain market environment, consider adding a loss-cutting mechanism (but be sure it’s not data mined)
Pitfall 5: Accounting for Outliers Improperly
Accounting for Outliers Improperly

Pitfall

- Removing outliers that are representative of the historical period studied
- Including outliers that are not representative of the historical period studied
- Removal or inclusion could bias the model, an effect that is particularly strong when outliers occur at the edge of the data (see Appendix)

Never miss an opportunity to study outliers


Les Iris, Vincent van Gogh
Some outliers should be left in...

**Example 1**

- The risk premium between the VIX and S&P 1m realized vol averages ~4 vol points since 2001
- Strategy: capture the volatility risk premium with a constant short vega position
- Chart below shows that realized vol can spike significantly above implied vol, leading to large drawdown
- Spikes recur, especially in times of uncertainty, and removal would favorably bias the strategy → don’t remove
...and some outliers should be removed

Example 2

- Consider the impact of quarterly changes in retail sales on 1-month stock volatility, for 30 retailers.
- Now, say one retailer announces a major strategy change, increasing subsequent realized vol by 10%.
- This outlier is due to a singular, non recurring event, and has a significant effect on the regression → remove
Identify outliers, and remove if appropriate

Solution

- Make sure to plot the data and visually identify possible outliers
- To gauge the relative influence of each outlier, recalculate the model with each outlier individually removed; see what changes occur

- Ask whether outliers are likely to recur in the future, and account for them in your strategy accordingly
  - Will Recur → include in backtest, and make necessary adjustments to strategy (e.g., Citi Dynavol)
  - Won’t Recur → remove from backtest, but point out and account for the removal

Out, Liar!

Your theory is wrong!

Source: http://davidmlane.com/ben/cartoons.html
Pitfall 6: Overfitting and Data Mining
Example 1

- Moving Averages Strategy: own S&P whenever 1wk MA > 2wk MA
- Backtest from 1990 until today: strategy outperforms S&P by ~600%age pts!

Source: Bloomberg.

“Buy S&P when 1wk MA > 2wk MA”
Strategy works, but how about 3wk MA > 4wk MA?

Example 1 (Cont’d)

- Now, own S&P whenever 3wk MA > 4wk MA
- Strategy underperforms
Trying too many strategies

Example 2

- The more strategies tested on the same dataset, the greater the risk that positive performance is due to a “lucky winner”
- A lucky winner strategy lacks predictive power out of sample

A Summer Analyst showed us a great long/short momentum algo

...but had tried 30 “strategies” in total
Example 3

- An investment manager gives market forecasts to $2^n x$ investors: he tells half (i.e. $2^{n-1}x$) that market will go up, and the other half that market will go down.
- Then, he drops the names to which he gave incorrect forecasts, and repeats the process for the remaining $2^{n-1}x$ investors.

After repeating this process $n$ times, the investment manager has given correct advice to $x$ investors.

These $x$ investors might be very impressed, unless aware of the number of incorrect forecasts given to others.

# Overfitting and Data Mining

## “Lucky Winners” Pitfall

- Testing many strategies until one outperforms
- Testing many variations of the same strategy until one outperforms

## Solutions

- Ask if small changes in trading rules cause big changes in performance
- Ask how many models were attempted, and what were their results
- If trigger is “vol > 2.34σ”, then the strategy is probably data mined
- Ask for cross validation results
- Extend time period (See MinBTL* in Appendix)

## “Too Many Parameters” Pitfall

- Too many parameters in a model improve “in sample fit”, but degrade “out of sample” performance

## Solutions

- Ask if the model was tested out-of-sample? (ie. Live date). Ask for cross validation results
- Perform statistical tests
  - T-test, F-test, Likelihood Ratio Test
  - Compare model Information Criterion
- Regularize models: Lasso, Ridge Regression

Pitfall 7: Overlooking Mark to Market Performance
Going Bankrupt through Martingale Betting

Pitfall

- Excellent P&L at the end, but you may go bankrupt on the way there

Example 1

- Coin-toss: bet $100, double bet after every “loss” (tails)
- Final Outcome: on first “win” (heads), recover all previous losses AND win original stake
- Or, go bankrupt before you win

Solution

- Run simulations of the strategy and look at mark-to-market performance
- Compute VaR and drawdowns
Insurance Companies and the Ultimate Ruin

Example 2

- Insurance companies have two opposing cash flows:
  - **Premiums** flow in at a constant rate
  - **Claims** arrive in a less predictable fashion

- According to statistical models, incoming premiums should cover claims, but there is still the probability of “ultimate ruin” if **too many claims come in at once**

Solution

- Classic actuarial solution: compute minimum capital required to avoid ruin with a certain % probability
A VIX Strategy

Example 3

- **Strategy**
  - Own S&P when VXX > 30
  - Don’t own when VXX < 20
- P&L is great year over year, but weekly mark-to-market shows you could be in for big losses

Solution

- Chart and assess the periodic mark-to-market performance
- Calculate max drawdown
- Calculate the Sharpe ratio

Source: Bloomberg
Pitfall 8: Failing to Expect the Unexpected

Source: http://www.despair.com/preservation.html
Failing to Expect the Unexpected

Pitfall

• Failing to account for a risk that has never happened yet

Examples

Not accounting for the possibility of:

• Implied correlations going above 1 in a liquidity crisis
• Decoupling among VIX exposure ETFs (ex: SVXY, XIV, VXX) during a surge in vol

Solution

• Stress-test the strategy to see how it performs in hypothetical crisis situations
• Build a loss-cutting mechanism into the strategy to protect against downside

Pitfall 9: Underestimating Hidden Exposures
Underestimating Hidden Exposures

Pitfall

• Thinking return is alpha when in reality, it’s beta

Examples

• Call overwriting strategy tested over sample period – what fraction of returns comes from variance risk premium vs. difference in delta combined with market movement?
• Combination strategies – combining of option greeks gives inaccurate view of risk (ex: vega impact on a calendar spread depends not only on level of vol but on term structure as well)

Solution

• Run strategy through a risk attribution model in order to gain a comprehensive understanding of factor exposures at play
• Talk to your Citi Structuring or Strategy desk
Pitfall 10: Forgetting Practical Aspects
Forgetting Practical Aspects

Pitfall

• Overlooking factors that affect whether a backtested strategy is feasible and can be implemented in real life
• If strategies are predictable, have large size, target illiquid names, or are uni-directional, they may have outsized market impact, or may be front-run

Solution

• Sufficient liquidity in underlying
• Account for transaction costs
• “Buy” or “sell” signals occur before transaction time, for example:
  – Don’t use closing prices as a signal for a trade on the close
  – Don’t use seasonally adjusted figures that are released well after the non-seasonal print
Using Options to Overcome Possible Strategy Pitfalls
Using Options to Overcome Possible Strategy Pitfalls

<table>
<thead>
<tr>
<th>Pitfall</th>
<th>Solution through Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark-to-Market Risk</td>
<td>A strategy may have positive expected return, but unacceptable mark-to-market risk. It may be possible to hedge this risk affordably through the use of specific derivatives strategies (ex: put-spreads, VIX-futures, etc.)</td>
</tr>
<tr>
<td>Failing to Expect Unexpected</td>
<td>A strategy may be overly exposed to tail risk (ex: currency risk from large movement in Swiss Franc) – possible to hedge with out-of-the-money options</td>
</tr>
<tr>
<td>Underestimating Hidden Exposures</td>
<td>Alpha strategy may have unwanted alternative beta exposure. Possible to net this out with options on Indices or ETFs</td>
</tr>
<tr>
<td>Forgetting Practical Aspects</td>
<td>A strategy may monetize a signal through illiquid securities, making it impractical – monetize instead through liquid options</td>
</tr>
</tbody>
</table>
Conclusion & Key Takeaways
1. **Never blindly trust a backtest**
   – Use the 10 pitfalls to test for robustness

2. **Understand Economic Rationale**
   – Understand what your backtest is trying to measure and the economic or structural rationale for why your strategy should work. Do you expect that this rationale will persist into the future?

3. **Talk to your backtester/structurer**
Questions to Ask the Structurer about a Backtest

Looking Beyond the Mean
• Did you calculate metrics other than the mean, such as st. dev., skewness, and Sharpe ratio?
• Can you show me a graph of the distribution?

Survivorship Bias
• Did you check the backtest for survivorship bias?
• If looking at constituents of an index, did you account for membership changes over time?

Correlation and Causation
• Does the alleged cause temporally precede the effect?
• Is predictive power strong?
• Did you conduct cross validation to test for biases that may be caused by a confounding variable?

Time Period
• What are the regimes included in your time period?
• Can I see the performance over each regime?
• Does the time period include a market environment similar to the current one?
• Does the time period include any extreme historical events?
• Does the strategy include a loss-cutting mechanism?
• Might the time period be too short / too long?

Outliers
• Did you identify possible outliers, and if so what are they?
• Did you assess the influence of each outlier?
• Do the outliers tell a story, or should they be removed?

Overfitting and Data Mining
• How sensitive is performance to changes in the parameters?
• How many models did you try before finding one that works?
• Can I see some statistical tests (T-test, F-Test, etc.)?
• Was the model tested out of sample?

Mark-to-Market
• Did you assess the strategy’s mark-to-market performance?
• Can I see the max drawdown?

Expecting the Unexpected
• What unexpected event or market environment could cause the strategy to blow up?

Hidden Exposures
• Did you run the strategy through a risk attribution model in order to scan for unidentified risk exposures?

Practical Aspects
• Can this strategy be arbitraged or front run?
• Did you account for transaction costs?
• If the signal occurs close to trading, what is the basis risk of using an earlier signal?
“I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”

-Enrico Fermi
Table of Contents

1. Outliers
2. Predictive Models
3. Overfitting
4. Cross Validation
Mathematical Appendix – Outliers

Methods to Detect Outliers

**Candlestick Plots:** A popular method to visually detect outliers is through a candlestick plot:

A candlestick plots the median, 25, and 75 percentile points, as well as the minimum and maximum points observed. If we notice a minimum or maximum that is extreme, this may indicate an outlier in the data.

**Standardized Residuals:** A more quantitative way to detect outliers in a regression setting is to investigate the distance of each data point from the final model result. In the examples above, it was clear that the outliers were much “further away” from the estimated regression line than any of the other data points. This notion may be quantified through the use of standardized residuals, which compute the normalized distance of each data point from our model prediction.

For each data point \( i \), the standardized Residual \( r(i) \) is computed as: \( r(i) = \frac{e_i}{\sqrt{\frac{1}{n} \sum e_i^2}} \), where \( e_i \) is the distance of our prediction from the actual realization. This residual is normalized, because the sample average of \( e_i \) is zero, and the sample variance is \( \frac{1}{n} \sum e_i^2 \).

With certain assumptions, we expect these residuals to be approximately normally distributed; therefore, any data points whose residuals have an absolute value greater than \( \pm 3 \) are possible outliers. These should be evaluated qualitatively to determine whether it makes sense to remove them. Some further refinements of this concept are Studentized and Jackknife residuals, which provide more accurate estimates of the variance of the residuals.
Mathematical Appendix – Outliers

Effects of Outliers on Model Results

All outliers matter, but those on the edges of the data matter more.

As an example, we consider 30 retailers that publish retail sales figures at the end of a given quarter. If we run a regression to model the impact of a change in a firm’s retail sales on the change in subsequent 1 month stock volatility, we can use the following formula:

\[ \Delta RVol_{1M} = \alpha + \beta (\Delta Sales_i) + \varepsilon_i \]

At right is a simulated regression for 30 retailers, with true \( \alpha = 0 \), \( \beta = -0.1 \), and \( \sigma = \frac{1}{2} \). We see that the estimated parameters, \( \hat{\alpha} = -0.11 \) and \( \hat{\beta} = -0.10 \) are close to the true model parameters.

\( \Delta RVol_{6M} \) is the change in 6 month realized volatility over the next 6 months
\( \Delta Sales_i \) is the change in retail sales over the last 6 months
\( \varepsilon_i \) is the normally distributed error (i.e. \( \varepsilon_i \sim N(0, \sigma) \))
Mathematical Appendix – Outliers

Effects of Outliers on Model Results

Imagine that one of our retailers announces a major strategy change, increasing uncertainty and subsequent realized volatility by an additional 10%. This increase is an exogenous shock not captured by our original model:

\[ \Delta RVol^{1M} = \alpha + \beta (\Delta Sales) + \epsilon + 10\% \]

This retailer is therefore an outlier, because it has a large idiosyncratic shock which does not exist in population at large. However, while any outlier may bias our results, the effect of outliers is most pronounced at the edges of our data:

An outlier located at the median of our data has a very small effect on our regression \( \hat{\beta} \) (it changes from -0.10 to -0.13), whereas the same exogenous shock has a profound effect on our results if it occurs at the extrema (our regression \( \hat{\beta} \) actually turns positive, changing to 0.01)
## Mathematical Appendix – Predictive Models

### Popular Predictive Models Used Today

There are many kinds of predictive models used today to build trading strategies. They include:

<table>
<thead>
<tr>
<th>Model</th>
<th>Idea</th>
<th>Variants &amp; Uses</th>
<th>Pros/Cons</th>
</tr>
</thead>
</table>
| **Linear Regression**  | One simple formula for computing predictions - each input contributes directly to the prediction through a linear $\beta$ weighting | **Variants:** Non-Linear Regression, Orthogonal Regression, Multi-Equation Systems, Generalized Least Squares, ARIMA($p,q$), ARCH/GARCH, Kalman Filter | **Pros:** Generally very intuitive, very flexible, easy to perform statistical tests, most popular methods  
**Cons:** Easy to datamine, cannot model complex non-linearities, not symmetrical |

\[
Returns_i = \alpha + \beta_1(CashFlow_i) + \beta_2(P/E_i) + \epsilon_i
\]

| **Support Vector Machines** | Model splits the data into two regions – data points which fall into a particular region are predicted to be of a particular type | **Variants:** Non-Linear Kernel Machines, Support Vector Regression | **Pros:** Easy to understand outputs - outputs match visual guesses, good predictive power  
**Cons:** Modelling non-linear data is more of an art than a science |

![Impact of P/E and Cash Flow on Earnings Expectation Performance](image-url)
### Popular Predictive Models Used Today (Cont’d)

<table>
<thead>
<tr>
<th>Model</th>
<th>Idea</th>
<th>Variants &amp; Uses</th>
<th>Pros/Cons</th>
</tr>
</thead>
</table>
| **Decision Trees**        | Model the predictive process as a series of questions—depending upon the answers assign a value | Variants: Boosted Trees, Random Forests                   | **Pros:** Can capture complex non-linear relationships, performs well on large datasets  
**Cons:** Stability & global optimum vs. intuitiveness & local optima trade-off |
|                           |                                                                       | Uses: Both classification and regression                 |                                                                          |
| **Artificial Neural Networks** | Model the predictive process as a network of neurons—based on inputs, the neurons will fire “yes” or “no” to produce an output value | Variants: Recurrent Neural Networks                       | **Pros:** Can capture very complex non-linear relationships               |
|                           |                                                                       | Uses: Both classification and regression                 | **Cons:** Completely un-intuitive, require diverse training data         |

#### Diagrams
- **Decision Trees**
  - P/E < 5
  - P/E > 5
  - CF < 5
  - CF > 5

- **Artificial Neural Networks**
  - Cash Flows
  - P/E
  - Input Layer
  - Hidden Layer
  - Output Layer
  - P/E > 5
  - P/E < 5
  - CF < 5
  - CF > 5
Minimum Backtest Length

The more models we test on the same dataset, the greater the risk that positive performance is due to a “lucky winner,” i.e. a model that performs well in-sample, but has no predictive power out-of-sample. One way to mitigate this risk is through the concept of Minimum Backtest Length (MinBTL). Here we assume that annualized Sharpe Ratios are random variables, which are approximately normally distributed.

MinBTL asks:

“Given that we have attempted $N$ independent strategies, and the best strategy achieved a Sharpe Ratio of $X$ in-sample, what is the number of years of backtest data needed to avoid selecting a strategy with an expected out-of-sample Sharpe Ratio of zero?”

An upper bound on MinBTL is given as:

$$\text{MinBTL} < \frac{2 \cdot \ln[N]}{X^2} \text{ in years}$$

Therefore, the Minimum Backtest Length forms a prudent rule of thumb for the minimum amount of backtest data needed - the more models we attempt, the further back we should look.

Mathematical Appendix – Overfitting From Too Many Signals

Statistical Model Selection

To prevent overfit, trading signals built into a model should be tested for significance:

**Practical Significance**: Does this signal materially impact my returns, Sharpe Ratio, Worst Drawdown, etc?

**Statistical Significance**: What is the probability that the effect observed arose purely by chance?

Statistical significance is usually evaluated through the computation of p-values: a p-value of 0.1 implies a 10% probability that the effect arose by chance. A p-value of 5% is the widely accepted cutoff for statistical significance.

To choose between different models and to test for statistical significance, the following concepts are often used:

<table>
<thead>
<tr>
<th>Concept</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-test/F-test</td>
<td>For regressions, test if one/all parameter(s) is/are statistically significant – both methods rely on the assumptions that errors are normally distributed</td>
</tr>
<tr>
<td>Likelihood Ratio Test</td>
<td>For non-linear regressions, compare the fit of two models, testing whether difference in fit is statistically significant</td>
</tr>
<tr>
<td>Cross Validation</td>
<td>Compare multiple models on simulated out-of-sample performance</td>
</tr>
<tr>
<td>Bayesian Information Criterion &amp; Akaike Information Criterion</td>
<td>Similar methods to choose between different models. Both reward good model fit (as measured by the log-likelihood function) but penalize model complexity</td>
</tr>
<tr>
<td>Model Regularization</td>
<td>Reduce the complexity of models; complexity penalty $\lambda | \beta |_x$ is added to the standard loss function, where the choice of penalty $\lambda$ is chosen through cross-validation, and norm $x$ is chosen depending on desired use (the 1-norm is used to reduce number of parameters in a model, the 2-norm is used to dampen extreme parameters). Examples: Lasso Regression (with 1-norm), Ridge Regression (with 2-norm)</td>
</tr>
</tbody>
</table>
Mathematical Appendix – Cross Validation

Methods of Cross Validation

One way to simulate the out-of-sample performance of a strategy is to partition the sample dataset into two parts: a new in-sample component on which the strategy is calibrated, and a new out-of-sample component on which we test performance. For example, if we are to test a model that PMI is a predictor of 6M future sales:

There are many popular variants of cross validation, each of which is conceptually similar. Some examples include:

- **Leave-One-Out**: for each data point in the sample, calibrate the model as though this point was out-of-sample, and assess the model’s predictive power over this left-out point
- **K-Fold**: cut the sample data into K continuous parts. For each part in the sample, calibrate the model as though that part was out-of-sample, and assess the model’s predictive power over this left-out part
Mathematical Appendix – Cross Validation

Methods of Cross Validation

Because successive calibrations over different data sets will yield differing model parameters, cross validation is effective at detecting non-robust strategies – the non-robust models will have inconsistent performance over the different out-of-sample periods. However, it is still possible to datamine cross validated strategies, especially if too many strategies are tested.

The primary metrics used to assess cross validation performance are:

- Trading strategy profitability, Sharpe ratio, worst drawdown, etc.
- Forecast mean squared error; i.e. the sample average of the squared error between the prediction and the actual occurrence: \( \frac{1}{n} \sum [(\hat{x} - x)^2] \), where \( \hat{x} \) is the predicted value and \( x \) is the actual occurrence.
References


- Hagin, Robert L., and Ronald N. Kahn. "What Practitioners Need to Know...About Backtesting."


- Harvey, Campbell R., and Liu, Yan. “...and the Cross-Section of Expected Returns.”

Disclaimer

In any instance where distribution of this communication is subject to the rules of the US Commodity Futures Trading Commission ("CFTC"), this communication constitutes an invitation to consider entering into a derivatives transaction under U.S. CFTC Regulations §§ 1.71 and 23.605, where applicable, but is not a binding offer to buy/sell any financial instrument.

This communication is issued by a member of the sales and trading department of Citigroup Global Markets Inc. or one of its affiliates (collectively, "Citi"). Sales and trading department personnel are not research analysts, and the information in this communication is not intended to constitute "research" as that term is defined by applicable regulations. Unless otherwise indicated, any reference to a research report or research recommendation is not intended to represent the whole report and is not in itself considered a recommendation or research report. All views, opinions and estimates expressed in this communication (i) may change without notice and (ii) may differ from those views, opinions and estimates held or expressed by Citi or other Citi personnel.

This communication is provided for information and discussion purposes only. Unless otherwise indicated, (i) it does not constitute an offer or recommendation to purchase or sell any financial instruments or other products, (ii) it does not constitute a solicitation if it is not subject to the rules of the CFTC (but see discussion above regarding communications subject to CFTC rules), and (iii) it is not intended as an official confirmation of any transaction. Unless otherwise expressly indicated, this communication does not take into account the investment objectives or financial situation of any particular person. Citi is not acting as an advisor, fiduciary or agent. Recipients of this communication should obtain advice based on their own individual circumstances from their own tax, financial, legal and other advisors about the risks and merits of any transaction before making an investment decision, and only make such decisions on the basis of the investor's own objectives, experience and resources. The information contained in this communication is based on generally available information and, although obtained from sources believed by Citi to be reliable, its accuracy and completeness cannot be assured, and such information may be incomplete or condensed. Any assumptions or information contained in this document constitute a judgment only as of the date of this document or on any specified dates and is subject to change without notice. Citi often acts as an issuer of financial instruments and other products, acts as a market maker and trades as principal in many different financial instruments and other products, and can be expected to perform or seek to perform investment banking and other services for the issuer of such financial instruments or other products.

The author of this communication may have discussed the information contained therein with others within or outside Citi and the author and/or such other Citi personnel may have already acted on the basis of this information (including by trading for Citi's proprietary accounts or communicating the information contained herein to other customers of Citi). Citi, Citi's personnel (including those with whom the author may have consulted in the preparation of this communication), and other customers of Citi may be long or short the financial instruments or other products referred to in this communication, may have acquired such positions at prices and market conditions that are no longer available, and may have interests different from or adverse to your interests.

Investments in financial instruments or other products carry significant risk, including the possible loss of the principal amount invested. Financial instruments or other products denominated in a foreign currency are subject to exchange rate fluctuations, which may have an adverse effect on the price or value of an investment in such products. This document does not purport to address all risks or material considerations which may be associated with entering into any transaction. Citi accepts no liability for any loss (whether direct, indirect or consequential) that may arise from any use of the information contained in or derived from this communication.

This document may contain historical and forward looking information. Past performance is not a guarantee or indication of future results. Any prices, values or estimates provided in this communication (other than those that are identified as being historical) are indicative only and may change without notice and do not represent firm quotes as to either price or size, nor reflect the value Citi may assign a security in its inventory. Forward looking information does not indicate a level at which Citi is prepared to do a trade and may not account for all relevant assumptions and future conditions. Actual conditions may vary substantially from estimates which could have a negative impact on the value of an instrument. You should contact your local representative directly if you are interested in buying or selling any financial instrument or other product or pursuing any trading strategy that may be mentioned in this communication. These materials are prepared solely for distribution into jurisdictions where such distribution is permitted by law. These materials are for the internal use of the intended recipients only and may contain information proprietary to Citi which may not be reproduced or redistributed in whole or in part without Citi's prior consent.

Although Citibank, N.A. (together with its subsidiaries and branches worldwide, "Citibank") is an affiliate of Citi, you should be aware that none of the financial instruments or other products mentioned in this communication (unless expressly stated otherwise) are (i) insured by the Federal Deposit Insurance Corporation or any other governmental authority, or (ii) deposits or other obligations of, or guaranteed by, Citibank or any other insured depository institution.

IRS Circular 230 Disclosure: Citi and its employees are not in the business of providing, and do not provide, tax or legal advice to any taxpayer outside of Citi. Any statements in this communication to tax matters were not intended or written to be used, and cannot be used or relied upon, by any taxpayer for the purpose of avoiding tax penalties. Any such taxpayer should seek advice based on the taxpayer’s particular circumstances from an independent tax advisor.

Options

This material may mention options. Before buying or selling options you should obtain and review the current version of the Options Clearing Corporation booklet, Characteristics and Risks of Standardized Options. A copy of the booklet can be obtained upon request from Citigroup Global Markets Inc., 390 Greenwich Street, 3rd Floor, New York, NY 10013 or by clicking the following link, http://www.theocc.com/about/publications/character-risks.jsp.

If you buy options, the maximum loss is the premium. If you sell put options, the risk is the entire notional below the strike. If you sell call options, the risk is unlimited. The actual profit or loss from any trade will depend on the price at which the trades are executed. The prices used herein are historical and may not be available when you order is entered. Commissions and other transaction costs are not considered in these examples. Option trades in general and these trades in particular may not be appropriate for every investor. Unless noted otherwise, the source of all graphs and tables in this report is Citi. Because of the importance of tax considerations to all option transactions, the investor considering options should consult with his/her tax advisor as to how their tax situation is affected by the outcome of contemplated options transactions.

© 2014 Citigroup Global Markets Inc. Member SIPC. All rights reserved. Citi and Citi and Arc Design are trademarks and service marks of Citigroup Inc. or its affiliates and are used and registered throughout the world.