FE570 Financial Markets and Trading
(Ref. Anatoly Schmidt CHAPTER 12 Back-Testing of Trading Strategies)

Steve Yang

Stevens Institute of Technology

12/11/2012
Overview of Back-Test of Trading Strategies "It is hard to make predictions, especially about the future" - Mark Twain.

- Forecasting has been widely used in economics, finance, and natural science (e.g., weather forecasting, climate change, global warming, outbreaks war, etc.) - arguably in every field where time series analysis is involved.

- The term back-testing implies that the forecasting models are fitted and tested using past empirical data. Usually, the entire available data set is split into two parts, one of which (earlier data) is used for in-sample calibration of the predicting model while the other is reserved for out-of-sample testing of the calibrated model. In simple models, it is usually assumed that the testing sample is variance-stationary; that is, the sample volatility is constant.

- In general, it is not always true that "more is better", as the long time series may be non-stationary.
Overview of Back-Test of Trading Strategies

- Very often we find that the optimal strategy parameters may evolve in time. In this case, moving-window sampling may be appropriate. For example, if a 10-year data sample is available, the first five-year data are used for in-sample calibration and the sixth-year data are used for out-of-sample testing. Then, the data from the second to the sixth year data are used for in-sample calibration and the seventh year is tested as out-of-sample data so on and so forth.

- Time time series my also experiences regime shifts caused by macro-economic events or changes in regulatory policies (e.g., introduction of the Euro in 1999, changes in uptick rule, etc.). Markov-switching models are sometimes used for a unified description of data samples with regime shifts.

- Lastly, there is always a danger of model over-fitting (too good to be true with regard to in-sample accuracy).
Overview of Back-Test of Trading Strategies

Figure: In this case, the model is fitted to the in-sample noise rather than to the deterministic relationship. The maximum likelihood-based criteria that have been used to evaluated the fitness of the model, such as, AIC, BIC, ABIC, DIC (Deviance information criterion), etc.

Figure 4.1: The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_{12}, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.
Overview of Back-Test of Trading Strategies

- **Data Snooping Bias** Generally, this bias can appear during the testing of different strategies with the same data set.
  
  A well-publicized example of data snooping is offered by Sullivan et al. (1999), who quoted Leinweber’s finding that the best predictor in the United Nations database for the S&P 500 stock index is the production of butter in Bangladesh.

- A general solution to the snooping bias is data *resampling*. Bootstrap is arguably the most popular resampling procedure. Other resampling technique *jackknife*, and *permutation* tests.

- The *Markov Chain Monte Carlo* (MCMC) simulation is another resampling approach, where the original sample is used to estimate probabilities of new returns conditioned on current and past returns.

- Finally, the *random entry protocol* is a simple resampling procedure that addresses the problem of correlation in two coupled time series.
Performance Measures for trading strategies usually relate to a rather long time period (at least a quarter or a year), during which a given strategy is used multiple times. Due to market factors, such as, finite liquidity, bid/ask spread, and transaction cost, practical performance measures should be evaluated after fees.

- Total return is the ultimate performance benchmark. It is calculated over some trading period after all long and short positions in the trading portfolio are closed. In other words, it is realized return that matters. Note that the compounded return usually listed in asset management statements is not realized and may include dividends and reinvestment. We are interested in pure trading strategy performance and therefore use the same notional amount in every round-trip trade.

- Percent of winning trades, \( p \), is another important performance measure.
Performance Measures:

- If this percentage and the ratio of the average winning amount to average losing amount, $r$, is assumed to be stable, one can use the Kelly’s criterion for estimating the optimal fraction of trading capital, $f$, to be used in each trade:

$$f = \frac{pr - 1 + p}{r}$$

It can be shown that the Kelly’s criterion is equivalent to choosing the trading size that maximizes the geometric mean of outcomes. Note that the Kelly’s formula yields an estimate that is valid only asymptotically. Therefore, risk-averse practitioners are advised to use a value of $f$ lower than the Kelly’s criterion suggests.

- The total number of trades for a given period is also important. Frequent trading may lead to more volatile outcomes.
**Performance Measures:**

- Multiple trades with a given strategy generate a probability distribution that can be used for hypothesis testing, in particular for comparing different strategies. Hence, the average return, \( \mu \), and its variance, \( \sigma \), are very important performance measures.

- Note that a positive return being accompanied with a high variance does not guarantee the strategy’s quality. Provided that the return distribution is normal, one can use the \( t \)-statistic for testing the hypothesis that the distribution mean is zero. Namely, for a given number of round-trip trades, \( N \), one calculates the \( t \)-value:

\[
t = \frac{\mu}{\left(\frac{\sigma^2}{N}\right)^{1/2}}
\]

then, the \( t \)-value can be used for finding statistical significance (also called \( p \)-value) from the Student’s distribution. Usually, the null hypothesis in an analysis of trading strategies is that the strategy return is zero.
Performance Measures:

- If two strategies have the same average return, the one with lower variance is more attractive. Provided that the return distributions are normal, two trading strategies $A$ and $B$ can be compared using the $t$-statistic:

  \[ t = \frac{\mu_A - \mu_B}{\left(\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}\right)^{1/2}} \]

  In this case, the degree of freedom equal $N_A + N_B - 2$. The $t$-statistic can also be used for analysis of profitability of a single strategy if indexes $A$ and $B$ refer to buy and sell signals.

- The Sharpe ratio is often used in performance analysis. Sometimes, the Sortino ratio is chosen instead. In the latter ratio, only negative returns are included in calculating the standard deviation $\sigma$. If a trading strategy performance is compared with performance of an index (or buy-and-hold strategy), the information ratio can be used.
Performance Measures:

- **Information ratio** is calculated as:

\[
IR = \frac{E[r_i] - E[r_0]}{\sigma_{i0}}
\]

where \( r_i \) is given return, \( r_0 \) is return of an index, and \( \sigma_{i0} \) is the tracking error (standard deviation between returns of the strategy and returns of the index).

- The **maximum drawdown** (MD) is another important risk measure, particularly for leveraged trades. For a process \( X(t) \) on the interval \( [0, T] \),

  In other words, MD is the largest drop of price after its peak.

\[
MD = \max\{\max(X[s] - X[t]) \}, \quad t \in [0, T], s \in [0, t]
\]

However, if drift can be neglected, expectation of MD has a simple analytic form:

\[
E[MD] = 1.2533\sigma \sqrt{T}
\]
Bootstrap

- The $t$-statistic can be misleading when it is applied to non-normal distributions. In the simple case, the bootstrap protocol is based on picking up at random an element of a given sample size $N$, copying it into the new sample, and putting it back (replacement). This random selection continues until the new sample has the same number of elements as the original one.

- Sometimes, blocks of several sequential elements of a given sample are picked at once (block bootstrap). Usually, the block bootstrap is implemented with replacement and blocks are not overlapping. Such an approach may preserve short-range autocorrelations present in the original sample. While a simple estimate of an optimal block size $L \sim N^{-1/3}$ can be used, choice of $L$ in the general case is not trivial. This method ensures the stationarity of samples bootstrapped from the stationary data - *stationary bootstrap*. 
Bootstrap

The block size $L$ in *stationary bootstrap* is randomly drawn from the geometric distribution:

$$Pr(L = k) = (1 - p)^{k-1}p$$

The average block size for the stationary bootstrap case equals $1/p$, which can serve as bridge between the size of the simple block bootstrap and the stationary bootstrap parameter $p$.

- A more sophisticated approach implies estimating a mathematical model that fits the given sample and bootstrapping the model’s residuals. Typical models used for stock prices are the random walk with drift, the AR models and the GARCH models.

- The number of bootstrapped samples needed for good accuracy may reach from several hundred to several thousand.
Markov Chain Monte Carlo

- Markov process is a generic stochastic process determined with relationships between its future, present, and past values.
- By definition, the Markov chain of the $k$th order is such a sequence of random variables $X_1, X_2, ..., X_n$ which satisfies the following equation:

\[
Pr(X_n = x | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_1 = x_1) = Pr(X_n = x | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_{n-k} = x_{n-k})
\]

In other words, only $k$ past values (sometimes called initial conditions) determine the present value. In particular, for $k = 1$, only one initial condition is needed:

\[
Pr(X_n = x | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_1 = x_1) = Pr(X_n = x | X_{n-1} = x_{n-1})
\]
Markov Chain Monte Carlo

- A Markov chain is stationary (or time-homogeneous) if probability in the left-hand side of the general form does not depend on index $n$.

- Generally, Markov variables can assume only a finite number of values (states). Stationary Markov chains of the 1st order with $N$ states are determined with $N^2$ probabilities $Pr(X_n = x_k | X_{n-1} = x_i) = p_{ik}$, $i, k = 1, 2, ..., N$. These probabilities are called the *transition kernel*, and the complete set of values $p_{ik}$ is called the *transition matrix*. Note that for each $k$

  $$\sum_{i=1}^{N} p_{ik} = 1$$

- Similary, the high order Markov chains of $k$th order are determined by $N^k$ propabilities.
Markov Chain Monte Carlo

- In MCMC-based resampling, the transition matrix is assumed stationary and is calculated using the original sample. Then, drawings from the uniform distribution are mapped onto transition probabilities for generating new samples.

- For example, consider a two-state Markov chain with \( p_{11} = p, p_{22} = q \) (which implies that \( p_{12} = 1 - p \) and \( p_{21} = 1 - q \)). Say the current state is 1. If a drawing from the uniform distribution is less than or equal to \( p \), then the next state is 1; otherwise, it is 2. If the current state is 2 and a drawing from the uniform distribution is less than or equal to \( q \), then the next state is 2; otherwise, it is 1.

- Since the transition matrix size grows with the Markov chain’s order as the power law, the use of higher orders for multi-state models can become a computational challenge. Normally, the financial returns do not have long memory, and low-order Markov chains should suffice for their resampling.
Random Entry Protocol

- In the general case, trading strategies can be determined not only by price dynamics but also by some liquidity measure(s), such as the bid/ask spread and the asset amount available at the best price, which also varies with time. Depending on the problem addressed with resampling, one may either want to preserve or destroy correlations between two time series.

- If correlations between two samples are weak, MCMC can be implemented for both coupled samples independently, and bootstrap for coupled samples can be reduced to picking up pairs of variables at the same time.

- The random entry protocol is choosing a random time within the given sample for each newly submitted order. It helps to avoid the bias due to autocorrelations in returns but preserves autocorrelations in order block size. It is similar to the stationary bootstrap approach.
Comparing Trading Strategies

As Brock et al. (1992) pointed out, "There is always the possibility that any satisfactory results obtained may simply be due to chance rather than to any merit inherent in the method yielding the results".

- White (2000) offered the Bootstrap Reality Check (BRC) for avoiding data snooping. The null hypothesis in BRC is that the performance of the best technical trading rule is no better than the performance of the benchmark.

- Hansen (2005) has shown that BRC may have a lower power due to the possible presence of poorly performing strategies in the test.

- Romnano and Wolf (2005) suggested that using a stepwise protocol to enhance the data snooping technique. This approach is focused on the family wide error (FWE) rate that is defined as the probability of rejecting at least one correct null hypothesis.