# Portfolio Backtesting

[Nematrian website page: PortfolioBacktesting, © Nematrian 2015]

## Abstract

Backtesting can be thought of as a short-hand way of seeking to working out whether some sort of forecasting approach might work in the future without actually having to wait for the future to arrive. In essence we develop algorithms that identify what results our models would have generated had we been running them at different points in time in the past, and we work out how well they would have subsequently performed. It is particularly important to be aware of the scope for 'look-back bias' in any such exercises. Backtesting can be applied to many different types of model in quantitative finance, including risk models and return forecasting models.

## Contents

## 1. Introduction

[PortfolioBacktesting1]

1.1 Portfolio backtesting comes in two main (related) forms:

(a) Backtesting the return generating potential of a particular investment strategy, and

(b) Backtesting the forecasting ability of a risk model.

1.2 In either case, backtesting can be thought of as a short-hand way of seeking working out whether some sort of forecasting approach might work in the future without actually having to wait for the future to arrive. In (a) we are forecasting, in effect, the *first moment of the distribution*, i.e. the mean drift of the relative return that might arise were we to follow a particular investment strategy. In (b) we are forecasting, in effect, second and higher moments, by testing the spread of returns that should have arisen in the past were the model to be accurate versus the spread of returns that actually did arise.

1.3 The aim of this and the following pages is to explore this topic further and to comment on the range of tools that can be used for such exercises. They build on material on backtesting (of risk models) contained in Kemp (2009). These tools need to be slightly more sophisticated than we might first expect, because in the past we would not have had the same amount of information as we have now.

1.4 We focus principally in these pages on backtesting risk models, because in some sense it is a more general mathematical problem than one focusing mainly on the first moment of the distribution, and because it also intimately relates to *calibration*, a topic that is explored in some detail in Kemp (2009).

## 2. Backtesting of risk models

[PortfolioBacktesting2]

2.1 One key reason why investors care about risk measurement (both the computation of *Value-at-Risk* type risk statistics and also the use of *stress tests* if derived in a similar manner), is that it provides a guide, albeit imperfect, regarding the potential range of future outcomes that the investor's portfolio might experience if invested in a particular way. This means that the risk models underlying the computations involved are amenable to verification by comparing predictions with actual future outcomes.

2.2 There are two ways of thinking about risk model *backtesting*:

(a) It can be thought of as a quick and 'cheap' way to carry out such a comparison without actually having to wait for the future to arrive. It involves identifying how well a risk model would have worked *in the past* had it been applied to the positions then present.

(b) It can also be thought of as a core step in the calibration of a (time series based) risk model to (past) market behaviour. To calibrate such a risk model to observed market behaviour, we parameterise the risk model in a suitable fashion and we choose which parameters to adopt by finding the model variant that best fits the data.

2.3 Backtesting also has a prominent (if sometimes just implicit) role in regulatory frameworks. Regulatory frameworks have increasingly incentivised firms to use their own risk models when determining their own regulatory capital requirements. Such models typically need to be approved by regulators before they can be used in such a manner. Given the complexity of the types of firms most likely to go down this route, it is not surprising that regulators are less than sanguine about their own ability to mitigate the possibility that firms might adopt overly optimistic assumptions in risk modelling. Hence, these regulatory frameworks also often include elements that penalise firms in capital terms if their risk models too often seem to underestimate actual magnitudes of outcomes. This makes it natural for firms to want to understand how well their risk models might have worked in the past (and for regulators to want to be provided with such information before approving a firm's model).

2.4 For firms opting to use industry-wide regulator-specified capital computations, backtesting might appear somewhat less important. However, this is because it has been (or ought to have been) carried out by the regulator itself when specifying the computation in question.

2.5 More generally, as risk measurement and management have acquired greater importance in business management it is natural for greater scrutiny to be placed on the validity of risk measures. Backtesting provides one way of 'quality assuring' such statistics.

## 3. In-sample versus out-of-sample backtesting

[PortfolioBacktesting3]

3.1 Short-cutting the future by referring merely to the past introduces *look-back bias*. Exactly how this works out in practice depends on how the backtesting is carried out.

3.2 One way of carrying out a backtest would be to take a single model of how the future might evolve and then to apply the *same* model to every prior period. This is called *in-sample* backtesting. The key issue with such an approach is that the model will typically have been formulated by reference to past history including the past that we are then testing the model against. Thus, unless we have been particularly inept at fitting the past when constructing the risk model in the first place, we should find that it is a reasonable fit in an in-sample, i.e. *ex-post*, comparison. We cannot then conclude much from its apparent goodness of fit.

3.3 Backtesters attempt to mitigate this problem by using so-called *out-of-sample* testing. What this involves is a specification of how to construct a model using data only available up to a particular point in time. We then apply the model construction algorithm *only* to observations that occurred *after* the end of the sample period used in the estimation of the model, i.e. out of the sample in question. The model might be estimated once-off using a particular earlier period of time and then the same model might be applied each time period thereafter. Alternatively, the model might be re-estimated at the start of each time period using data that would have then been available, so that the time period then just about occur is still (just) after the in-sample period.

3.4 Whilst out-of-sample modelling does reduce look-back bias it does not eliminate it. Risk models ultimately involve lots of different assumptions about how the future might evolve, not least the format of the risk model itself. In the background there are lots of competing risk models that we might have considered suitable for the problem. Not too surprisingly, the only ones that actually see the light of day, and therefore get formally assessed in an out-of-sample context, are ones that are likely to be tolerably good at fitting the past even in an out-of-sample context. Risk modellers are clever enough to winnow out ones that will obviously fail such a test before the test is actually carried out. This point is perhaps more relevant to backtesting of return generating algorithms, given the human tendency to rationalise explanations for success or failure, perhaps even if there is no such explanation, see e.g. Taleb (2004).

## 4. Testing backtest quality statistically

[Nematrian website page: PortfolioBacktesting4, © Nematrian 2015]

### Aggregate quality

[PortfolioBacktesting4a]

4.1 Any statistic such as a VaR estimate that is ultimately derived in part from analysis of a finite data sample is itself just an uncertain estimate of whatever is its ‘true’ underlying (but ultimately unobservable) value. It therefore comes with some error. Moreover, outcomes that arise in the future will also ultimately be probabilistic in nature.

4.2 Thus, suppose we experienced a significantly adverse outcome in the next period, well outside the typical spread of ranges we might have otherwise predicted. Does this mean that our model is wrong? Not necessarily. It might just mean that we have been unlucky.

4.3 Statisticians face this sort of issue with any type of modelling. The way that it is typically tackled is to postulate a hypothesis and to then identify the likelihood that the hypothesis is wrong (with the model being rejected if the hypothesis is too likely to be wrong). But even then, we might have alighted on the right model but might reject it because of a fluke series of outcomes.

4.4 Statistical backtesting of risk models typically thus proceeds in one of two ways:

(a) We tabulate past estimates from our risk model (with suitable out-of-sample adjustments as appropriate) of the specific statistic that we are most interested in 'estimating correctly' versus past outcomes. For example, the statistic in question might be a given quantile level, i.e. a suitable VaR estimate. We then apply suitable statistical tests applicable to that particular statistic, see e.g. Campbell (2006), Hurlin and Tokpavi (2006), Pena, Rivera and Ruiz-Mata (2006) or Zumbach (2006) to test if past actuals suitably fit past predictions. For example, we might use a one sided likelihood ratio test which provides a confidence interval on the number of rejects that we would expect to see, rejecting the model if too many actuals exceed corresponding predictions.

(b) Alternatively, we may seek to test whether the entire distributional form that our model would have predicted when applied to past data seems to fit the observed range of actual past outcomes, using appropriate statistical tests, see e.g. Campbell (2006) or Dowd (2006).

4.5 Statistical techniques might also typically be supplemented by corresponding graphical comparison of the data. This might, for example, indicate visually that the model was a poor fit only during a limited 'exceptional' period in the past which might permit some suitable explanation or refinement of the model to cater for this historic period.

## Fitting 'period by 'period'

[PortfolioBacktesting4b]

4.6 Commonly, we want the model not only to fit the data in aggregate but also to fit it 'period by period'. By this we mean that we want exceptionally adverse outcomes to occur apparently randomly through time rather than being strongly clumped together into narrow time windows. The latter might imperil the solvency of a firm more than the former, since there would be less time during such a window to generate new profits or raise new capital needed to maintain a solvent status or credible business model.

4.7 Campbell (2006) explains that the problem of determining whether a 'hit' sequence (i.e. for, say, VaR, an indicator of the form $I_t(\alpha)$ which is 1 if the actual outcome for time period $t$ is worse than the $\alpha$-quantile VaR, and 0 otherwise) is acceptable involves two key properties, namely:

(a) *unconditional coverage*, i.e. actual probability of occurrence when averaged through time should match expected probability of occurrence; and

(b) *independence*, i.e. that any two elements of the hit sequence should be independent of each other.

4.8 The former can be tested for by using, for example, Kupiec's (1995) test statistic as described in Campbell (2006), which involves a proportion of failures $POF$, defined as follows, where there are $T$ observations:

$$POF = 2\log\left(\left(\frac{1-\hat{\alpha}}{1-\alpha}\right)^{T-I(\alpha)}\left(\frac{\hat{\alpha}}{\alpha}\right)^{I(\alpha)}\right)$$

where $\hat{\alpha} = \frac{1}{T} I(\alpha)$ = observed number of failures, $I(\alpha) = \sum_{t=1}^{T} I_t(\alpha)$

4.9 Alternatively it can be tested for by using a $z$-statistic also described in Campbell (2006):

$$z = \frac{\sqrt{T}(\hat{\alpha} - \alpha)}{\sqrt{\alpha(1-\alpha)}}$$

4.10 Campbell (2006) also describes several ways of testing for independence, including Chrisftofferson's (1998) Markov test (which examines whether the likelihood of a VaR violation at time $t$ depends on whether or not a VaR violation occurred at time $t-h$ by building up a contingency table). This idea could presumably be extended to correlations between times further apart. He also describes a more recent test suggested by Christofferson and Pelletier (2004) which uses the insight that if VaR violations are independent of each other then the amount of time between them should also be independent, which hristofferson and Pelletier apparently argue may be a more powerful test than the Markov test. Campbell (2006) also describes ways of testing for unconditional coverage and independence simultaneously.

## Nomenclature

[PortfolioBacktestingNomenclature]

$\alpha$ = confidence level  
$\hat{\alpha}$ = observed number of failures  
$I_t(\alpha)$ = failure indicator (at $\alpha$ confidence level)  
$POF$ = proportion of failures

## References

[PortfolioBacktestingRefs]

Campbell, S. D. (2006). A review of backtesting and backtesting procedures. *Journal of Risk*, **9**, No. 2, pp. 1-17

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, 39, pp. 841-62

Christoffersen, P. and Pelletier, D. (2004). Backtesting value-at-risk: a duration-based approach. *Journal of Empirical Finance*, **2**, pp. 84-108

Dowd, K. (2006). Backtesting market risk models in a standard normality framework. *Journal of Risk*, **9**, No. 2, pp. 93-111

Hurlin, C. and Tokpavi, S. (2006). Backtesting value-at-risk accuracy: a simple new test. *Journal of Risk*, **9**, No. 2, pp. 19-37

Kemp, M.H.D. (2009). *Market consistency: Model calibration in imperfect markets*. John Wiley & Sons [for further information on this book please see Market Consistency]

Pena, V. H. de la, Rivera, R. Ruiz-Mata, J. (2006). Quality control of risk measures: backtesting VAR models. *Journal of Risk*, **9**, No 2, pp. 39-54

Zumbach, G. (2006). Backtesting risk methodologies from one day to one year. *Journal of Risk*, **9**, No 2, pp. 55-91