RiskMetrics Group™ Research Monthly

Back to Backtesting

Christopher C. Finger
chris.finger@riskmetrics.com

May 2005

Nine years ago, the Basel Committee released the Market Risk Amendment (MRA) to the 1988 Capital Accord. The MRA allows banks to assess the risk of their trading operation, and whether the bank has adequate capital to cover this risk, using an internal risk model. While the MRA is flexible regarding the specifics of the internal models themselves, it is quite specific as to how the internal model should be validated, and to how failures of this validation should affect the bank's required capital. Validation, or backtesting, was consequently an active research area in the mid-1990s.

In Europe, the UCITS[1] directive, passed in the late 1980s, provides for harmonized registration and compliance standards for fund managers. It establishes a European "passport," whereby as long as a fund is certified in one EU country, it may be marketed in the rest of the EU. Under the initial directive, funds were restricted by compliance guidelines, and their use of derivatives was for the most part limited to hedging activities. An amended directive, UCITS 3, passed in 2002, admits position-taking in derivatives and relaxes other investment restrictions, provided the fund can demonstrate a risk management process that monitors the fund's exposure and that can be used to communicate with the appropriate regulatory bodies. The individual EU countries are tasked with interpreting UCITS 3 and in particular, defining an appropriate risk management process.

BaFin, the German financial supervisory authority, published its interpretation of UCITS 3 in its Derivative Regulation circular of February 2004. In the circular, BaFin introduces a "qualified" method of risk assessment. Under this method, fund managers may invest in derivatives, provided they use a recognized risk model and limit the Value-at-Risk (VaR) of the fund to no more than twice the VaR of the fund's benchmark. Similarly to the MRA, the BaFin circular leaves the details of the risk model largely up to the fund manager, provided the manager can demonstrate appropriate understanding and documentation. Also similarly to the MRA, the circular requires that the manager backtest the risk model, comparing actual results to model forecasts. The BaFin circular, however, is noticeably less specific regarding the implications of the backtesting, and does not specify the consequences of bad results. Rather, regarding backtesting failures, it states simply that "the Authority must be kept regularly informed of this anomaly, its magnitude, and the reason it arose."

While the BaFin language may sound ominous, it actually represents a commendable step forward from the MRA's approach to backtesting. Rather than prescribing a method and set of consequences to be applied to each institution, the BaFin circular pushes the responsibility of innovation to the institution, and requires in return that the institution be transparent about its methods and open with its model results.

[1] Undertakings for Collective Investment in Transferable Securities

At the same time, regulation in other areas, notably pensions and insurance,[2] is also moving toward the use of internal risk models, raising the issue of validation and backtesting there as well. As a result, backtesting is an open topic again, meaning it is an appropriate time to look forward to new innovation, as well as to review some of the issues that dominated the dialogue on backtesting nine years ago.

In this note, we examine two issues, broadly what to test and how to test, as well as the backtesting experience of a number of banks. We argue that under the UCITS 3 framework, asset managers are well positioned to improve on the backtesting legacy left by the Market Risk Amendment.

What to test

Simply put, backtesting involves comparing ex ante risk forecasts to ex post realizations of the portfolio profit-and-loss (P&L), with the aim of identifying whether the risk model is performing well. From a statistician's point of view, it is tempting to jump directly to a discussion of which tests are appropriate to perform. However, the most critical issue is the seemingly mundane matter of which P&L to actually compare to the risk measure.

In most discussions of backtesting, two types of P&L are defined: Actual and Hypothetical. The Actual P&L is quite simple, and includes all gains and losses from market moves, trading revenue and fee income. The Hypothetical P&L is the P&L that would have resulted if the portfolio had stayed constant over the period in question; thus, it excludes both trading revenue and fee income. The MRA suggests that banks develop the capability to test against either notion, though in practice, tests against only the Actual P&L are acceptable in many countries.

For our purposes, we consider two types of Hypothetical P&L: Market and Model. In both cases, we assume the portfolio stays constant. For the Market Hypothetical P&L, we consider the actual market valuation changes for each instrument in the portfolio.[3] For the Model Hypothetical P&L, we compute valuation changes using the same functions as exist in the risk model. To clarify the distinction, consider two examples.

Approximate pricing functions. Suppose for some options, a risk model does not include a full option valuation, but rather approximates the change in option value by the option delta times the change in the option underlying. In this case, the Model Hypothetical P&L for the option is just the option delta multiplied by the actual change in the option underlying; the Market Hypothetical P&L is the actual change in option value. This case applies as well to complex fixed income securities whose values are approximated by sensitivities to key interest rate points.

Non-modeled risk factors. Suppose for some corporate bonds, the risk model incorporates changes in the base yield curve, but assumes the bond's specific spread to the base curve is constant. Again, the Market Hypothetical P&L is the actual change in the bond price. The Model Hypothetical P&L is

[2] The Financial Assessment framework for Dutch pension funds and the Swiss Solvency Test for private insurance firms are recent examples.

[3] Note here that if we close a position during one day, then to the Market Hypothetical P&L requires a valuation of the position at a time we no longer hold it. This is likely a mere inconvenience with exchange-traded securities, but could be a significant practical barrier with over-the-counter positions.

obtained through the bond pricing function by applying the realized change in the base curve, but not the change in spread. A similar example is an option for which vega risk (that is, the risk from changes in the implied volatility) is not modeled. In general, this category applies to most examples of *specific risk*: the risk of an instrument over and above its sensitivity to common market risk factors.

A discussion of which of the three notions of P&L is correct is somewhat misguided. In fact, comparing our risk forecasts to any of these provides us with useful information.

Examining the Model Hypothetical P&L provides for a clean test of our assumptions about the evolution of risk factors. We know under this approach that if we forecast the risk factors precisely, we will exactly match the P&L; any poor backtesting results are reason to question our assumptions about risk factor evolution, such as our estimates of volatility and correlation. This backtest does not, however, inform us on the completeness of the risk model. If our forecasts are good for what risk factors we do model, but we either ignore relevant factors (such as spreads or implied volatilities) or apply inadequate pricing approximations, we will still see good backtest performance.

Comparing our risk forecasts to the Market Hypothetical P&L combines a test of our risk factor forecasts with a test of our pricing models. If we apply this comparison after concluding (based on the previous test) that our risk factor forecasts are accurate, then this test of the mark-to-market valuations informs us on the completeness of our pricing models. A failure of this type of test, after a successful Model Hypothetical test, indicates that either we have not modeled a relevant risk factor or that we have utilized inadequate approximations to the true pricing functions. We see then that our distinction between the two Hypothetical P&L's isolates our assumptions about how risk factors evolve from those about how risk factor changes translate into portfolio P&L.

A comparison with the Actual P&L is the ultimate test of the model's performance. It is not advisable, however, to perform this test in isolation, since a failure is difficult to interpret. A failure here could indicate that the volatility of intraday trading revenue or fee income overwhelms the risk of the constant holdings. However, a failure could also derive from poor risk factor forecasts, incomplete risk factor coverage or inadequate pricing models, possibilities that can be eliminated through testing with the Hypothetical P&L. In fact, once we establish that our model accurately forecasts the Hypothetical P&L, our next step should be a simple comparison of the Market Hypothetical and Actual P&L. If these two series are dramatically different, then we know any risk model that assumes a constant portfolio will not, by construction, forecast the Actual P&L well. It is possible that such a difficulty could be resolved simply by shortening our risk horizon, but most likely, we should consider other modeling approaches that somehow incorporate the trading and fee revenues.

Berkowitz and O'Brien (2002) examine Actual P&L and VaR forecasts for six US banks, and apply a time series model directly to the P&L data. They find that in most cases, their time series model outperforms the banks' VaR forecasts. Partly, this is an indication that the VaR forecasts are themselves deficient; more crucially, though, this result illustrates the extent to which the actual bank P&L is driven by factors that are not modeled. The results also raise the question of whether it would be sensible to abandon the VaR models altogether, and rely on the P&L time

series approach. For only assessing portfolio VaR, this might be sensible, but portfolio models do enable more than just a portfolio VaR calculation; stress testing and what-if analysis, for instance, would not be possible under the time series approach.

How to test

The simplest test of a VaR model is to count the number of days on which the realized portfolio loss is greater than the VaR forecast. We define such days as VaR exceptions. The proportion of VaR exceptions should be consistent with the stated VaR confidence level. This is referred to as a test of *unconditional coverage*, and is at the heart of the backtesting framework stipulated by the MRA. Under the MRA's "traffic light" approach, banks are required to count the number of days over the prior year (250 trading days) on which the portfolio loss exceeded the 99% VaR forecast. With zero to four such exceptions, the backtest is qualified as in the green zone; with five to nine exceptions, in the yellow zone; and with ten or more exceptions, in the red zone. The MRA describes regulatory responses appropriate to each zone.

A common criticism of tests of unconditional VaR coverage is their lack of power, that is, their ability to differentiate statistically a good VaR model from a bad one. Under the MRA framework, if the 99% VaR forecasts are accurate, we would expect the portfolio loss to exceed VaR on 2.5 of the 250 trading days. However, even if the model is accurate, and if we assume that the profit and loss relative to the VaR is independent from one day to the next (which is an assumption we will revisit later), then there is still a good chance (over five percent) that five or more of the 250 days are exceptions. Thus, to be relatively certain that we not reject a good model, we must admit the statistical possibility that there are more exceptions than we expect. By admitting this, however, we open ourselves to the possibility that we fail to identify a bad model.

Consider the MRA yellow zone, where we reject any model that produces five or more exceptions. This leaves us with a five percent chance of rejecting a good model. Suppose we also have a flawed model which produces a VaR coverage level of just 98%. Even with this flawed model, there is a 45% chance that four or fewer exceptions occur over 250 days. Thus, there is an almost even chance that we accept a model under which the true expected number of exceptions is twice our target. Even extending the historical data is of little help here: with four years (1000 days) of data, there is still a 10% chance we accept the flawed model if we fix the 5% chance of rejecting the good one.

With the red zone (ten or more exceptions), there is a different tradeoff. Here, there is only a 0.01% chance that we reject a good model, but a 97% chance that we accept the same flawed model. With 1000 days of data, the same probability of rejecting the good model corresponds to a still large 80% chance of accepting the flawed version.

A formerly common mistake with testing for unconditional coverage was to simply compute the average VaR level over the testing period, and to count the number of days on which the realized loss exceeds this average level.[4] While a test of this form is appealing for its simplicity, it is flawed as long as the portfolio VaR is not constant throughout the testing period.

[4] A determined reader can find such a test in some bank annual reports from the late 1990s.

Table 1: Example: backtesting using average VaR

	Risky	Safe	Total
Days	190	60	250
99% VaR	10	100	31.6
Exp. days w/ Loss > 10	1.9	24.5	26.4
Exp. days w/ Loss > 31.6	0.0	13.9	13.9
Exp. days w/ Loss > 100	0.0	0.6	0.6

Consider an example. Assume that we have a good VaR model, and that our portfolio is normally distributed. Suppose that for the first nine months (190 days) of the year, our portfolio is relatively safe, with a VaR of 10 every day. For the last three months (60 days), with bonus day approaching, our traders stretch for greater profits, and the VaR is 100 every day. (See Table 1.) Overall, the average portfolio VaR is 31.6. During the safe period, we expect 1.9 VaR exceptions (days on which the loss exceeds 10); we also expect the portfolio loss to never[5] exceed the average VaR of 31.6. During the risky period, we expect the loss to exceed 31.6 on 13.9 days, and to exceed the VaR level (100) on 0.6 days. Overall, we expect losses to exceed the average VaR level (31.6) on 13.9 days; in other words, we expect losses to exceed our average VaR on over five percent of the trading days. The naive conclusion might be to reject our perfectly good VaR model.

So what has happened? By not comparing realized losses to the VaR for each specific day, we have erroneously identified exceptions. During the risky period, we expect about 13 days where the loss falls between the average VaR (31.6) and the true VaR (100). These days are not exceptions, and yet we have identified them as such. On the other hand, during the safe period, we expect 1.9 days on which the loss falls between the true VaR (10) and the average VaR (31.6), and we have not identified these as exceptions. So not only have we produced an inflated count of VaR exceptions, we have also failed to identify some days on which exceptions did truly occur.

In addition to illustrating why this simple VaR averaging test is flawed, the prior example introduces the concept of *conditional coverage*. Testing for conditional coverage involves testing not just for how many exceptions occur but also for when they occur. In the example above, we assumed that our model did provide appropriate conditional coverage: the model could distinguish between a safe day and a risky day. For such a model, the likelihood that the next day produces a VaR exception should not be influenced by how volatile the market is, what positions are in the portfolio, or whether an exception occurred recently; all such information should be embedded in the current VaR forecast. The implication is that the timing of VaR exceptions should be uniform; VaR exceptions should not occur in clusters. Poor conditional coverage can arise with VaR models that do

[5] The probability, under the normal distribution, of a return being greater than three times the 99% VaR is about one in 10^{13}.

not adjust quickly enough to rising market volatility, resulting in a tendency to produce sequences of consecutive exceptions. As well, poor conditional coverage can arise when models overadjust, reacting to a single market event with a much higher VaR forecast, and producing a tendency for one exception to not be followed by another exception for some time.

Christoffersen (1998) presents a statistical test for conditional coverage alone, in which he tests for whether an exception on one day influences the likelihood of an exception on the next, as well as a joint test for unconditional and conditional coverage. Other tests of conditional coverage could attempt to uncover whether particular market conditions or portfolio holdings tend to coincide with more exceptions.

With only information about the frequency and timing of VaR exceptions, tests for conditional and unconditional coverage represent the limit of possible backtesting schemes. Further tests involve examination of the entire forecasted P&L distribution or of the magnitude of VaR exceptions. These require more information about the VaR model, since the expected results depend not only on the VaR confidence level, but on the assumed distribution.

Bank experience under the MRA

Enough time has now passed to examine some bank experience under the MRA framework. Berkowitz and O'Brien (2002) examine P&L and VaR data for six large US banks over the period 1998-2000. The P&L data represents the Actual P&L, including trading revenue and fee income. Overall, the banks' VaR forecasts appear quite conservative: only one bank shows exceptions on more than 1% of trading days, one bank shows no exceptions at all,[6] and the overall proportion of exceptions is less than 0.5%. Furthermore, most of the exceptions occur during the three months encompassing the Russian default crisis of 1998. Not surprisingly, this clustering results in two of the banks failing tests of conditional coverage.

In addition to reporting P&L and VaR to their regulators, a number of banks also provide such information in their annual reports to shareholders. We examined annual reports for three banks which have reported VaR and backtesting results: JPMorgan Chase, Deutsche Bank and Société Générale.

JPMorgan Chase reports backtesting results using Actual P&L. They report no VaR exceptions in 2004, two in 2003, none in 2002, and one in 2001. The VaR results are certainly conservative, as the bank states. In fact, the probability that so few exceptions occur over four years, assuming that the VaR model is accurate, is less than one percent: so few exceptions are inconsistent with a true VaR confidence level of 99%. Similarly, Société Générale reports backtesting using Actual P&L. They report no exceptions over the period 2002-2004. The probability of so few exceptions occurring is about 0.04%. Deutsche Bank also reports no exceptions over 2002-2004 using Actual P&L. Deutsche Bank differs from the other two banks, though, in that it discusses backtesting against Hypothetical P&L as well.

The Deutsche Bank comments serve as a reminder that the German regulators have been the most insistent on backtesting using Hypothetical P&L. The study by Jaschke, Stahl and Stehle (2003), while on the surface comparable to that of Berkowitz and O'Brien (2002), differs most significantly in that it

[6] The probability, assuming that the VaR model is good, of observing no exceptions over such a history is about 0.3%.

uses Hypothetical, rather than Actual P&L; consequently, the results are not clouded by trading revenue and fee income. The study study focuses on daily P&L and VaR data for thirteen German banks in 2001. In contrast to the US results, the number of exceptions is consistent with the 99% VaR confidence level, and the authors conclude that the VaR models of the considered banks are "essentially okay."

So while some backtesting schemes offer disappointing results, the observations with Hypothetical P&L suggest that the risk models are doing what they are supposed to do, but being compared to data they were never meant to forecast. Anticipating the application of backtesting to funds gives us cause for optimism, since the fee and trading income is likely to be less important than the actual market value changes on the portfolio positions. We are hopeful, then, that the backtesting exercises for funds under the new regulatory environment are at least more relevant comparisons of P&L and VaR data.

What next?

Looking ahead, we return to the BaFin circular, where the emphasis is not on zones of exceptions, but on communication between the fund and its regulator, specifically regarding the causes for VaR exceptions. Thus, it is crucial that we think not just of tests for the accuracy of VaR models, but also of diagnostic tools that allow us to explain exceptions when they occur.

Broadly, we can consider two questions about a VaR exception: first, which part of the portfolio or market was at the root of the exception, and second, which aspect of the model failed. To the first point, disaggregating the portfolio along common dimensions (geography, asset classes, maturity buckets, etc.) and performing backtests on the individual subportfolios can inform us as to whether an exception is a localized or more global event. To the second point, backtesting against the three notions of P&L mentioned previously lets us evaluate in turn whether an exception is related to inadequate risk factor forecasts, coarse pricing approximations, or unmodeled portfolio effects.

Further reading

- Basel Committee on Banking Supervision (1996). Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements, Technical Note.
- Berkowitz, J. and O'Brien, J. (2002). How accurate are value-at-risk models at commercial banks?, *Journal of Finance*, **57**(3): 1093–1112.
- Christoffersen, P. (1998). Evaluating interval forecasts, *International Economic Review*, **39**: 841–862.
- Duncan, J. (2004). Backtesting for asset managers: An overview and practical implications, Presentation, RiskMetrics Group European Client Conference.
- Jaschke, S., Stahl, G., and Stehle, R. (2003). Evaluating VaR forecasts under stress – the German experience, Center for Financial Studies, Working Paper 2003/32.