Methods for Evaluating Value-at-Risk Estimates

Jose A. Lopez

I. Current Regulatory Framework

In August 1996, the U.S. bank regulatory agencies adopted the market risk amendment (MRA) to the 1988 Basle Capital Accord. The MRA, which became effective in January 1998, requires that commercial banks with significant trading activities set aside capital to cover the market risk exposure in their trading accounts. (For further details on the market risk amendment, see *Federal Register* [1996].) The market risk capital requirements are to be based on the value-at-risk (VaR) estimates generated by the banks' own risk management models.

In general, such risk management, or VaR, models forecast the distributions of future portfolio returns. To fix notation, let y_t denote the log of portfolio value at time t. The k-period-ahead portfolio return is $\varepsilon_{t+k} = y_{t+k} - y_t$. Conditional on the information available at time t, ε_{t+k} is a random variable with distribution f_{t+k}. Thus, VaR model m is characterized by f_{mt+k}, its forecast of the distribution of the k-period-ahead portfolio return.

VaR estimates are the most common type of forecast generated by VaR models. A VaR estimate is simply a specified quantile (or critical value) of the forecasted f_{mt+k}. The VaR estimate at time t derived from model m for a k-period-ahead return, denoted $VaR_{mt}(k,\alpha)$, is the critical value that corresponds to the lower α percent tail of f_{mt+k}. In other words, VaR estimates are forecasts of the maximum portfolio loss that could occur over a given holding period with a specified confidence level.

Under the "internal models" approach embodied in the MRA, regulatory capital against market risk exposure is based on VaR estimates generated by banks' own VaR models using the standardizing parameters of a ten-day holding period ($k = 10$) and 99 percent coverage ($\alpha = 1$). A bank's market risk capital charge is thus based on its own estimate of the potential loss that would not be exceeded with 1 percent certainty over the subsequent two-week period. The market risk capital that bank m must hold for time $t+1$, denoted MCR_{mt+1}, is set as the larger of $VaR_{mt}(10,1)$ or a multiple of the average of the previous sixty $VaR_{mt}(10,1)$ estimates, that is,

$$MRC_{mt+1} = max\left[VaR_{mt}(10,1);\ S_{mt} \times \frac{1}{60}\sum_{i=0}^{59} VaR_{mt-i}(10,1)\right] + SR_{mt},$$

where S_{mt} is a multiplication factor and SR_{mt} is an additional capital charge for the portfolio's idiosyncratic credit risk. Note that under the current framework $S_{mt} \geq 3$.

The S_{mt} multiplier explicitly links the accuracy of a bank's VaR model to its capital charge by varying over time. S_{mt} is set according to the accuracy of model m's VaR

Jose A. Lopez, formerly an economist at the Federal Reserve Bank of New York, is now an economist at the Federal Reserve Bank of San Francisco.

estimates for a one-day holding period ($k = 1$) and 99 percent coverage, denoted $VaR_{mt}(1,1)$ or simply VaR_{mt}. S_{mt} is a step function that depends on the number of exceptions (that is, occasions when the portfolio return ε_{t+1} is less than VaR_{mt}) observed over the last 250 trading days. The possible number of exceptions is divided into three zones. Within the green zone of four or fewer exceptions, a VaR model is deemed "acceptably accurate," and S_{mt} remains at its minimum value of three. Within the yellow zone of five to nine exceptions, S_{mt} increases incrementally with the number of exceptions. Within the red zone of ten or more exceptions, the VaR model is deemed to be "inaccurate," and S_{mt} increases to its maximum value of four.

II. Alternative Evaluation Methods

Given the obvious importance of VaR estimates to banks and now their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. To date, two hypothesis-testing methods for evaluating VaR estimates have been proposed: the binomial method, currently the quantitative standard embodied in the MRA, and the interval forecast method proposed by Christoffersen (forthcoming). For these tests, the null hypothesis is that the VaR estimates in question exhibit a specified property characteristic of accurate VaR estimates. If the null hypothesis is rejected, the VaR estimates do not exhibit the specified property, and the underlying VaR model can be said to be "inaccurate." If the null hypothesis is not rejected, then the model can be said to be "acceptably accurate."

However, for these evaluation methods, as with any hypothesis test, a key issue is their statistical power, that is, their ability to reject the null hypothesis when it is incorrect. If the hypothesis tests exhibit low power, then the probability of misclassifying an inaccurate VaR model as "acceptably accurate" will be high. This paper examines the power of these tests within the context of a simulation exercise.

In addition, an alternative evaluation method that is not based on a hypothesis-testing framework, but instead uses standard forecast evaluation techniques, is proposed. That is, the accuracy of VaR estimates is gauged by how well they minimize a loss function that represents the regulators' concerns. Although statistical power is not relevant for this evaluation method, the related issues of comparative accuracy and model misclassification are examined within the context of a simulation exercise. The simulation results are presented below, after the three evaluation methods are described. (See Lopez [1998] for a more complete discussion.)

Evaluation of VaR Estimates Based on the Binomial Distribution

Under the MRA, banks will report their VaR estimates to their regulators, who observe when actual portfolio losses exceed these estimates. As discussed by Kupiec (1995), assuming that the VaR estimates are accurate, such exceptions can be modeled as independent draws from a binomial distribution with a probability of occurrence equal to 1 percent. Accurate VaR estimates should exhibit the property that their unconditional coverage $\alpha^* = x/250$, where x is the number of exceptions, equals 1 percent. Since the probability of observing x exceptions in a sample of size 250 under the null hypothesis is

$$Pr(x) = \binom{250}{x} 0.01^x \times 0.99^{250-x},$$

the appropriate likelihood ratio statistic for testing whether $\alpha^* = 0.01$ is

$$LR_{uc} = 2[\log(\alpha^{*x}(1-\alpha^*)^{250-x}) - \log(0.01^x \times 0.99^{250-x})].$$

Note that the LR_{uc} test is uniformly most powerful for a given sample size and that the statistic has an asymptotic $\chi^2(1)$ distribution.

Evaluation of VaR Estimates Using the Interval Forecast Method

VaR estimates are also interval forecasts of the lower 1 percent tail of f_{t+1}, the one-step-ahead return distribution. Interval forecasts can be evaluated conditionally or unconditionally, that is, with or without reference to the information available at each point in time. The LR_{uc} test is an unconditional test since it simply counts exceptions over the entire period. However, in the presence of variance dynamics, the conditional accuracy of interval forecasts is an

important issue. Interval forecasts that ignore variance dynamics may have correct unconditional coverage, but at any given time, they will have incorrect conditional coverage. In such cases, the LR_{uc} test is of limited use since it will classify inaccurate VaR estimates as "acceptably accurate."

The LR_{cc} test, adapted from the more general test proposed by Christoffersen (forthcoming), is a test of correct conditional coverage. Given a set of VaR estimates, the indicator variable I_{mt+1} is constructed as

$$I_{mt+1} = \begin{cases} 1 \ if \ \varepsilon_{t+1} < VaR_{mt} \\ 0 \ if \ \varepsilon_{t+1} \geq VaR_{mt} \end{cases}.$$

Since accurate VaR estimates exhibit the property of correct conditional coverage, the I_{mt+1} series must exhibit both correct unconditional coverage and serial independence. The LR_{cc} test is a joint test of these two properties. The relevant test statistic is $LR_{cc} = LR_{uc} + LR_{ind}$, which is asymptotically distributed $\chi^2(2)$. The LR_{ind} statistic is the likelihood ratio statistic for the null hypothesis of serial independence against the alternative of first-order Markov dependence.

Evaluation of VaR Estimates Using Regulatory Loss Functions

The loss function evaluation method proposed here is not based on a hypothesis-testing framework, but rather on assigning to VaR estimates a numerical score that reflects specific regulatory concerns. Although this method forgoes the benefits of statistical inference, it provides a measure of relative performance that can be used to monitor the performance of VaR estimates.

To use this method, the regulatory concerns of interest must be translated into a loss function. The general form of these loss functions is

$$C_{mt+1} = \begin{cases} f(\varepsilon_{t+1}, VaR_{mt}) \ if \ \varepsilon_{t+1} < VaR_{mt} \\ g(\varepsilon_{t+1}, VaR_{mt}) \ if \ \varepsilon_{t+1} \geq VaR_{mt} \end{cases},$$

where $f(x,y)$ and $g(x,y)$ are functions such that $f(x,y) \geq g(x,y)$ for a given y. The numerical scores are constructed with a negative orientation, that is, lower values of C_{mt+1} are preferred since exceptions are given higher scores than nonexceptions. Numerical scores are generated for individual VaR estimates, and the score for the complete regulatory sample is

$$C_m = \sum_{i=1}^{250} C_{mt+i}.$$

Under very general conditions, accurate VaR estimates will generate the lowest possible numerical score. Once a loss function is defined and C_m is calculated, a benchmark can be constructed and used to evaluate the performance of a set of VAR estimates. Although many regulatory loss functions can be constructed, two are described below (see diagram).

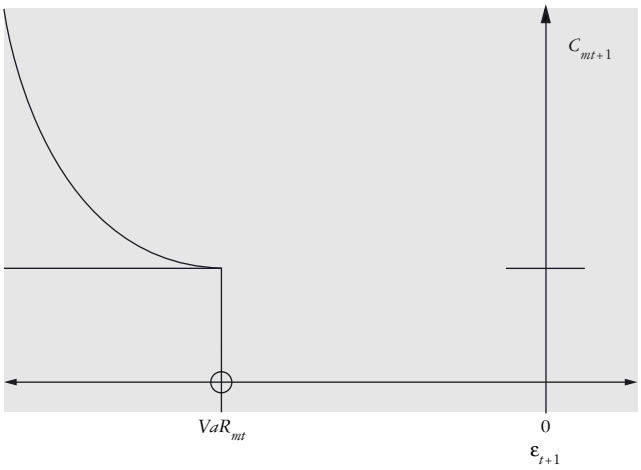
Loss Function Implied by the Binomial Method

The loss function implied by the binomial method is

$$C_{mt+1} = \begin{cases} 1 \ if \ \varepsilon_{t+1} < VaR_{mt} \\ 0 \ if \ \varepsilon_{t+1} \geq VaR_{mt} \end{cases}.$$

Note that the appropriate benchmark is the expected value of C_{mt+1}, which is $E[C_{mt+1}] = 0.01$, and for the full sample, $E[C_m] = 2.5$. As before, only the number of exceptions is of interest, and the same information contained in the binomial method is included in this loss function.

Loss Functions of Interest



Notes: The diagram graphs both the binomial and the magnitude loss functions. The binomial loss function is equal to 1 for $\varepsilon_{t+1} < VaR_{mt}$ and zero otherwise. For the magnitude loss function, a quadratic term is added to the binomial loss function for $\varepsilon_{t+1} < VaR_{mt}$.

Loss Function That Addresses the Magnitude of the Exceptions

As noted by the Basle Committee on Banking Supervision (1996), the magnitude as well as the number of exceptions are a matter of regulatory concern. This concern can be readily incorporated into a loss function by introducing a magnitude term. Although several are possible, a quadratic term is used here, such that

$$C_{mt+1} = \begin{cases} 1 + (\varepsilon_{t+1} - VaR_{mt})^2 & if \ \varepsilon_{t+1} < VaR_{mt} \\ 0 & if \ \varepsilon_{t+1} \geq VaR_{mt} \end{cases}.$$

Thus, as before, a score of one is imposed when an exception occurs, but now, an additional term based on its magnitude is included. The numerical score increases with the magnitude of the exception and can provide additional information on how the underlying VaR model forecasts the lower tail of the underlying f_{t+1} distribution. Unfortunately, the benchmark based on the expected value of C_{mt+1} cannot be determined easily, because the f_{t+1} distribution is unknown. However, a simple, operational benchmark can be constructed and is discussed in Section III.

Simulation Exercise

To analyze the ability of the three evaluation methods to gauge the accuracy of VaR estimates and thus avoid VaR model misclassification, a simulation exercise is conducted. For the two hypothesis-testing methods, this amounts to analyzing the power of the statistical tests, that is, determining the probability with which the tests reject the null hypothesis when it is incorrect. With respect to the loss function method, its ability to evaluate VaR estimates is gauged by how frequently the numerical score for VaR estimates generated from the true data-generating process (DGP) is lower than the score for VaR estimates from alternative models. If the method is capable of distinguishing between these scores, then the degree of VaR model misclassification will be low.

In the simulation exercise, the portfolio value y_{t+1} is specified as $y_{t+1} = y_t + \varepsilon_{t+1}$, where the portfolio return ε_{t+1} is generated by a GARCH(1,1)-normal process. That is, h_{t+1}, the variance of ε_{t+1}, has dynamics of the form $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$. The true DGP is one of eight VaR models evaluated and is designated as the "true" model, or model 1.

The next three alternative models are homoskedastic VaR models. Model 2 is simply the standard normal distribution, and model 3 is the normal distribution with a variance of 1½. Model 4 is the *t*-distribution with six degrees of freedom, which has fatter tails than the normal distribution and an unconditional variance of 1½.

The next three models are heteroskedastic VaR models. For models 5 and 6, the underlying distribution is the normal distribution, and h_{mt+1} evolves over time as an exponentially weighted moving average of past squared returns, that is,

$$h_{mt+1} = (1-\lambda)\sum_{i=0}^{\infty} \lambda^i \varepsilon_{t-i}^2 = \lambda h_{mt} + (1-\lambda)\varepsilon_t^2.$$

This type of VaR model, which is used in the well-known RiskMetrics calculations (see J.P. Morgan [1996]), is calibrated here by setting λ equal to 0.94 and 0.99 for models 5 and 6, respectively. Model 7 has the same variance dynamics as the true model, but instead of using the normal distribution, it uses the *t*-distribution with six degrees of freedom. Model 8 is the VaR model based on historical simulation using 500 observations, that is, using the past 500 observed returns, the α percent VaR estimate is observation number $5*\alpha$ of the sorted returns.

In the table, panel A presents the power analysis of the hypothesis-testing methods. The simulation results indicate that the hypothesis-testing methods can have relatively low power and thus a relatively high probability of misclassifying inaccurate VaR estimates as "acceptably accurate." Specifically, the tests have low power against the calibrated normal models (models 5 and 6) since their smoothed variances are quite similar to the true GARCH variances. The power against the homoskedastic alternatives is quite low as well.

For the proposed loss function method, the simulation results indicate that the degree of model misclassification generally mirrors that of the other methods, that is, this method has a low-to-moderate ability to distinguish between the true and alternative VaR models. However, in certain cases, it provides additional useful information on

SIMULATION RESULTS FOR GARCH(1,1)-NORMAL DGP
Units: percent

	Models						
	Homoskedastic			Heteroskedastic			Historical
	2	3	4	5	6	7	8
PANEL A: POWER OF THE LR_{UC} AND LR_{CC} AGAINST ALTERNATIVE VAR MODELS[a]							
LR_{uc}	52.3	21.4	30.5	5.1	10.3	81.7	23.2
LR_{cc}	56.3	25.4	38.4	6.7	11.9	91.6	33.1
PANEL B: ACCURACY OF VAR ESTIMATES USING REGULATORY LOSS FUNCTIONS[b]							
Loss function							
Binomial	91.7	41.3	18.1	52.2	48.9	0	38.0
Magnitude	96.5	56.1	29.1	75.3	69.4	0	51.5

Notes: The results are based on 1,000 simulations. Model 1 is the true data-generating process, $\varepsilon_{t+1}|\Omega_t \sim N(0,h_{t+1})$, where $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$. Models 2, 3, and 4 are the homoskedastic models $N(0, 1)$, $N(0,1.5)$, and t (6), respectively. Models 5 and 6 are the two calibrated heteroskedastic models with the normal distribution, and model 7 is a GARCH(1,1) -t (6) model with the same parameter values as model 1. Model 8 is the historical simulation model based on the previous 500 observations.

[a]The size of the tests is set at 5 percent using finite-sample critical values.

[b]Each row represents the percentage of simulations for which the alternative VaR estimates have a higher numerical score than the "true" model, that is, the percentage of the simulations for which the alternative VaR estimates are correctly classified as inaccurate.

the accuracy of the VaR estimates under the defined loss function. For example, note that the magnitude loss function is relatively more correct in classifying VaR estimates than the binomial loss function. This result is not surprising given that it incorporates the additional information on the magnitude of the exceptions into the evaluation. The ability to use such additional information, as well as the flexibility with respect to the specification of the loss function, makes a reasonable case for the use of the loss function method in the regulatory evaluation of VaR estimates.

III. IMPLEMENTATION OF THE LOSS FUNCTION METHOD

Under the current regulatory framework, regulators observe $\{\varepsilon_{t+i}, VaR_{mt+i}\}_{i=1}^{250}$ for bank m and thus can construct, under the magnitude loss function, C_m. However, for a realized value C_m^*, aside from the number of exceptions, not much inference on the performance of these VaR estimates is available. It is unknown whether C_m^* is a "high" or "low" number.

To create a comparative benchmark, the distribution of C_m, which is a random variable due to the random observed portfolio returns, can be constructed. Since each observation has its own distribution, additional assumptions must be imposed in order to analyze $f(C_m)$, the distribution of C_m. Specifically, the observed returns can be assumed to be independent and identically distributed (iid); that is, $\varepsilon_{t+1} \sim f$. This is quite a strong assumption, especially given the heteroskedasticity often found in financial time series. However, the small sample size of 250 mandated by the MRA allows few other choices.

Having made the assumption that the observed returns are iid, their empirical distribution $\hat{f}(\varepsilon_{t+1})$ can be estimated parametrically, that is, a specific distributional form is assumed, and the necessary parameters are estimated from the available data. For example, if the returns are assumed to be normally distributed with zero mean, the variance can be estimated such that $\hat{f}(\varepsilon_{t+1})$ is $N(0,\hat{\sigma}^2)$.

Once $\hat{f}(\varepsilon_{t+1})$ has been determined, the empirical distribution of the numerical score C_m under the distributional assumptions, denoted $\hat{f}(C_m)$, can be generated since the distribution of the observed returns and the corresponding VaR estimates are now available. For example, if $\varepsilon_{t+1} \sim N(0,\hat{\sigma}^2)$, then the corresponding VaR estimates are $VaR_{\hat{f}t} = -2.32\hat{\sigma}$. Using this information, $\hat{f}(C_m)$ can then be constructed via simulation by forming 1,000 values of the numerical score C_m, each based on 250 draws from $\hat{f}(\varepsilon_{t+1})$ and the corresponding VaR estimates.

Once $\hat{f}(C_m)$ has been generated, the empirical quantile $\hat{q}_m = \hat{F}(C_m^*)$, where $\hat{F}(C_m)$ is the cumulative distribution function of $\hat{f}(C_m)$, can be calculated for the observed value C_m^*. This empirical quantile provides a performance benchmark, based on the distributional assumptions, that can be incorporated into the evaluation of the underlying VaR estimates. In order to make this benchmark operational, the regulator should select a threshold quantile above which concerns regarding the performance of the VaR estimates are raised. This decision should be based both on the regulators' preferences and the severity of the distributional assumptions used. If $\hat{q}_m$ is below the threshold that regulators believe is appropriate, say, below 80 percent, then

C_m^* is "typical" under both the assumptions on $\hat{f}(\varepsilon_{t+1})$ and the regulators' preferences. If $\hat{q}_m$ is above the threshold, then C_m^* can be considered atypical, and the regulators should take a closer look at the underlying VaR model.

Note that this method for evaluating VaR estimates does not replace the hypothesis-testing methods, but instead provides complementary information, especially regarding the magnitude of the exceptions. In addition, the flexibility of this method permits many other concerns to be incorporated into the analysis via the choice of the loss function.

IV. Conclusion

As implemented in the United States, the market risk amendment to the Basle Capital Accord requires that commercial banks with significant trading activity provide their regulators with VaR estimates from their own internal models. The VaR estimates will be used to determine the banks' market risk capital requirements. This development clearly indicates the importance of evaluating the accuracy of VaR estimates from a regulatory perspective.

The binomial and interval forecast evaluation methods are based on a hypothesis-testing framework and are used to test the null hypothesis that the reported VaR estimates are "acceptably accurate," where accuracy is defined by the test conducted. As shown in the simulation exercise, the power of these tests can be low against reasonable alternative VaR models. This result does not negate their usefulness, but it does indicate that the inference drawn from this analysis has limitations.

The proposed loss function method is based on assigning numerical scores to the performance of the VaR estimates under a loss function that reflects the concerns of the regulators. As shown in the simulation exercise, this method can provide additional useful information on the accuracy of the VaR estimates. Furthermore, it allows the evaluation to be tailored to specific interests that regulators may have, such as the magnitude of the observed exceptions. Since these methods provide complementary information, all three could be useful in the regulatory evaluation of VaR estimates.

References

Basle Committee on Banking Supervision. 1996. "Supervisory Framework for the Use of 'Backtesting' in Conjunction with the Internal Models Approach to Market Risk Capital Requirements." Manuscript, Bank for International Settlements.

Christoffersen, P. F. Forthcoming. "Evaluating Interval Forecasts." International Economic Review.

Federal Register. 1996. "Risk-Based Capital Standards: Market Risk." Vol. 61: 47357-78.

J.P. Morgan. 1996. RiskMetrics Technical Document. 4th ed. New York: J.P. Morgan.

Kupiec, P. 1995. "Techniques for Verifying the Accuracy of Risk Measurement Models." Journal of Derivatives 3: 73-84.

Lopez, J. A. 1998. "Methods for Evaluating Value-at-Risk Estimates." Federal Reserve Bank of New York Research Paper no. 9802.