Regulatory Evaluation of Value-at-Risk Models

Jose A. Lopez

Economic Research Department
Federal Reserve Bank of San Francisco
101 Market Street
San Francisco, CA 94105-1579
(415) 977-3894
jose.a.lopez@sf.frb.org

Draft date: June 30, 1999

ABSTRACT:

Beginning in 1998, U.S. commercial banks may determine their regulatory capital requirements for financial market risk exposure using value-at-risk (VaR) models. Currently, regulators have available three hypothesis-testing methods for evaluating the accuracy of VaR models: the binomial, interval forecast and distribution forecast methods. Given the low power often exhibited by their corresponding hypothesis tests, these methods can often misclassify forecasts from inaccurate models as acceptably accurate. An alternative evaluation method using loss functions based on probability forecasts is proposed. Simulation results indicate that this method is only as capable of differentiating between forecasts from accurate and inaccurate models as the other methods. However, its ability to directly incorporate regulatory loss functions into model evaluations make it a useful complement to the current regulatory evaluation of VaR models.

JEL Primary Field Name: C52, G2
Key Words: value-at-risk, volatility modeling, probability forecasting, bank regulation

Acknowledgments: The views expressed here are those of the author and not necessarily those of the Federal Reserve Bank of San Francisco or the Federal Reserve System. I thank Philippe Jorion (editor of the *Journal of Risk*), Peter Christoffersen, Frank Diebold, Darryl Hendricks, Beverly Hirtle, Paul Kupiec, Jim O'Brien and Philip Strahan as well as seminar participants at the 1996 meeting of the Federal Reserve System Committee on Financial Structure and Regulation, the Wharton Financial Institutions Center Conference on Risk Management in Banking and the 1997 Federal Reserve Bank of Chicago Conference on Bank Structure and Competition for their comments.

Regulatory Evaluation of Value-at-Risk Models

ABSTRACT:

Beginning in 1998, U.S. commercial banks may determine their regulatory capital requirements for financial market risk exposure using value-at-risk (VaR) models. Currently, regulators have available three hypothesis-testing methods for evaluating the accuracy of VaR models: the binomial, interval forecast and distribution forecast methods. Given the low power often exhibited by their corresponding hypothesis tests, these methods can often misclassify forecasts from inaccurate models as acceptably accurate. An alternative evaluation method using loss functions based on probability forecasts is proposed. Simulation results indicate that this method is only as capable of differentiating between forecasts from accurate and inaccurate models as the other methods. However, its ability to directly incorporate regulatory loss functions into model evaluations make it a useful complement to the current regulatory evaluation of VaR models.

> My discussion of risk measurement issues suggests that disclosure of quantitative measures of market risk, such as value-at-risk, is enlightening only when accompanied by a thorough discussion of how the risk measures were calculated and how they related to actual performance. (Greenspan, 1996a)

I. Introduction

The profits of financial institutions are directly or indirectly tied to the behavior of financial time series, such interest rates, exchange rates and stock prices. This exposure is commonly referred to as "market risk". Over the past decade, financial institutions have significantly increased their use of econometric models to manage their market risk exposure for a number of reasons, such as their increased trading activities, their increased emphasis on risk-adjusted returns on capital and advances in both the theoretical and empirical finance literature. Given these developments, financial regulators have also begun to focus their attention on the use of such models by regulated institutions.

The main example of such regulatory concern is the 1996 amendment to the Basle Capital Accord, which requires that commercial banks with significant trading activities set aside capital to cover the market risk exposure in their trading accounts. The U.S. bank regulatory agencies adopted this amendment and began enforcing it in 1998. [1] Under the amended capital rules, banks' market risk capital charges can be based on the "value-at-risk" (VaR) estimates generated by their own VaR models. In general, such models forecast the time-varying distributions of portfolio returns, and VaR estimates are forecasts of the maximum portfolio loss that could occur over a given holding period with a specified confidence level; that is, a VaR estimate is a specified lower quantile of a forecasted distribution of portfolio returns.

Given the importance of VaR estimates to banks and now to their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. Three evaluation methods based on hypothesis tests have been proposed to date. In each of these tests, the null hypothesis is that the VaR forecasts in question exhibit a specified property characteristic of accurate VaR

[1] For a thorough discussion of the 1988 Basle Capital Accord and the U.S. implementation of the 1996 market risk amendment, see Wagster (1996) and Federal Register (1996), respectively. For a related discussion on the regulatory capital requirements for securities firms, see Dimsom and Marsh (1995).

forecasts. Specifically, the evaluation method based on the binomial distribution, currently the quantitative standard embodied in the 1996 amendment and extensively discussed by Kupiec (1995), examines whether VaR estimates, on average, provide correct coverage of the lower α percent tails of the forecasted distributions. The interval forecast method proposed by Christoffersen (1998) examines whether VaR estimates exhibit correct coverage at each point in time, and the distribution forecast method proposed by Crnkovic and Drachman (1996) examines whether empirical quantiles derived from a VaR model's distribution forecasts are independent and uniformly distributed. In these tests, if the null hypothesis is rejected, the VaR forecasts do not exhibit the specified property, and the underlying VaR model is 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 power; i.e., their ability to reject the null hypothesis when it is incorrect. If a hypothesis test exhibits poor power properties, then the probability of misclassifying an inaccurate VaR model as acceptably accurate will be high. This paper examines the power of these three tests within the context of a simulation exercise using several data generating processes.

In addition, this paper proposes an evaluation method based on the probability forecasting framework presented by Lopez (1997). In contrast to those listed above, this method uses standard forecast evaluation techniques and gauges the accuracy of VaR models by how well their probability forecasts minimize a loss function directly relevant to the user. By incorporating regulatory loss functions directly into the evaluation of VaR models, this method provides information on the performance of VaR models with respect to regulatory criteria, as opposed to the statistical criteria implied in the other methods. In this paper, the probability forecasts of interest are of specified regulatory events, and the loss function used is the quadratic probability score (QPS), whose value ranges over the interval [0,2]. VaR models with lower QPS values can be said to be more accurate than others.

A drawback of this evaluation method is that the properties of the QPS value for a particular model and specified event cannot be easily determined a priori, as opposed to the properties of the three aforementioned test statistics. Thus, this method cannot be used, as the other methods, to statistically test whether a VaR model is "acceptably accurate" or "inaccurate"

in an absolute sense. Instead, this method can be used to provide relative comparisons of model accuracy over different time periods and in relation to other VaR models, which should be useful, additional information for model users in general and regulators in particular. This method's ability to address the issues of VaR model misclassification and comparative accuracy under different loss functions is also examined within the context of a simulation exercise.

The simulation results indicate that the hypothesis-testing methods can have relatively low power and thus a relatively high chance of misclassifying an inaccurate VaR model as "acceptably accurate". With respect to the probability forecasting method, the simulation results indicate that the QPS values for the accurate VaR models are less than those for the inaccurate models a high percentage of the time. Further analysis, using hypothesis-testing techniques that permit a power comparison across all four evaluation methods, indicates that this method's power is roughly in line with that of the other three methods. Thus, even though the proposed method is only as capable of differentiating between VaR models as the other methods, its ability to directly incorporate regulatory loss functions into model evaluations make it a useful complement to the statistical methods currently used in the regulatory evaluation of VaR models.

The paper is organized as follows. Section II describes both the current regulatory framework for evaluating VaR models and the four evaluation methods examined. Sections III and IV outline the simulation exercise and present the results, respectively. Section V concludes.

II. Evaluating VaR Models

VaR models are characterized by their forecasted distributions of k-period-ahead portfolio returns. To fix notation, let Y_t represent portfolio value at time t in dollar terms, and $y_t = \ln(Y_t)$. The k-period-ahead portfolio return is denoted $\varepsilon_{t+k} = y_{t+k} - y_t$. Conditional on the information available at time t, ε_{t+k} is a random variable with distribution f_{t+k}; that is, $\varepsilon_{t+k} \mid \Omega_t \sim f_{t+k}$. Thus, VaR model m is characterized by f^m_{t+k}, its forecast of f_{t+k}.

Currently, VaR estimates are the most common type of forecast generated from VaR models. A VaR estimate is a forecast of the maximum portfolio loss that could occur over a given holding period with a specified confidence level. The VaR estimate at time t derived from model m for a k-period-ahead return with α percent confidence, denoted $VaR^m_t(k,\alpha)$, is the

quantile of f_{t+k}^{m} that corresponds to its lower α percent tail. Thus, $VaR_t^m(k,\alpha)$ is the solution to

$$\int_{-\infty}^{VaR_t^m(k,\alpha)} f_{t+k}^{m}(x)\,dx = \frac{\alpha}{100},$$

or, equivalently, $VaR_t^m(k,\alpha) = F_{t+k}^{m\,-1}(\alpha/100)$, where F_{t+k}^{m} is the forecasted cumulative distribution function. Note that a VaR estimate is expressed in dollar terms as the difference between the current portfolio value and the portfolio value corresponding to it; that is, $VaR_t^m(k,\alpha)$ is expressed in dollar terms as $Y_t\left(1 - e^{VaR_t^m(k,\alpha)}\right)$.

Given their role in bank risk management and now in regulatory capital calculations, the evaluation of VaR estimates and the models underlying them is of interest to both banks and their regulators. Note, however, that the regulatory evaluation of such models differs from institutional evaluations in three important ways.[2] First, a regulatory evaluation has the goal of assuring that sufficient capital is available to protect an institution from significant portfolio losses, a goal that may not be shared by an institutional evaluation due to issues of moral hazard. Second, regulators, although potentially privy to the details of an institution's VaR model, generally cannot evaluate every component of the model and its implementation as well as the originating institution can. Third, regulators have the responsibility of constructing evaluations that are comparable across institutions. Thus, although individual banks and regulators may use similar evaluation methods, the regulatory evaluation of VaR models has certain unique characteristics that should be addressed.

In this section, the current regulatory framework for calculating market risk capital charges is described, and four methods for the regulatory evaluation of VaR models are discussed. The first three methods are based on testing the null hypothesis that the VaR forecasts in question exhibit specified properties characteristic of accurate VaR forecasts. The proposed fourth method is instead based on standard forecast evaluation techniques; that is, the relative accuracy of a VaR model is gauged by how well a specified regulatory loss function is minimized by the model's probability forecasts.

---

[2] For a general discussion of the differences between financial institutions and their regulators on the issues of risk measurement and capital allocation, see Estrella (1995).

A. Current Regulatory Framework

The current risk-based capital rules for the market risk exposure of large, U.S. commercial banks, effective as of 1998, are explicitly based on VaR estimates. The capital rules cover all assets in a bank's trading account (i.e., assets carried at their current market value) as well as all foreign exchange and commodity positions wherever located. Any bank or bank holding company whose trading activity accounts for more than ten percent of its total assets or is more than $1 billion must hold regulatory capital against their market risk exposure.

These capital charges are based on the VaR estimates generated by banks' own VaR models using the standardizing parameters of a ten-day holding period (k = 10) and 99 percent coverage (α = 1). In other words, a bank's market risk capital charge is based on its forecast of the potential portfolio loss that would not be exceeded over the subsequent two week period with one percent probability. The market risk capital that bank m must hold for time t+1, MRC_{t+1}^{m}, is set as the larger of the dollar value of $VaR_t^m(10,1)$ or a multiple of the average of the previous sixty $VaR_t^m(10,1)$ estimates in dollar terms; that is,

$$MRC_{t+1}^{m} = \max\left[Y_t\left(1 - e^{VaR_t^m(10,1)}\right);\ S_t^m * \frac{1}{60}\sum_{i=0}^{59} Y_{t-i}\left(1 - e^{VaR_{t-i}^m(10,1)}\right)\right] + SR_t^m,$$

where S_t^m and SR_t^m are a multiplication factor and an additional capital charge for the portfolio's idiosyncratic credit risk, respectively.[3] Note that, under the current framework, $S_t^m \geq 3$.

The S_t^m multiplier is included in the calculation of MRC_{t+1}^{m} for two reasons. First, as described by Hendricks and Hirtle (1997), it adjusts the specified VaR estimates to what regulators consider to be a minimum capital requirement that reflects their concerns regarding both prudent capital standards and model accuracy.[4] Second, S_t^m is used to explicitly link the accuracy of a bank's VaR model to its capital charge. In the current regulatory framework, S_t^m

[3] The specific risk capital charge is used to cover possible adverse price changes due to unanticipated, idiosyncratic events, such as an unexpected bond default. Although an important topic, specific risk is not examined here.

[4] Stahl (1997) provides a theoretical justification of the use of a regulatory multiplication factor. Using Chebyshev's inequality, he shows that a multiplication factor approximately equal to three can be used to account for the possible misspecification of the distribution underlying VaR estimates for α=1. Thus, S_t^m can be viewed as a regulatory adjustment for model error.

is set according to the accuracy of model m's VaR estimates for a one-day holding period (k = 1) and 99 percent coverage level (α = 1), denoted as $VaR_t^m(1,1)$.

S_t^m is a step function that depends on the number of exceptions -- defined as occasions when $\varepsilon_{t+1} < VaR_t^m(1,1)$ -- 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 through 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. The institution must also explicitly improve its risk management system.[5]

Since capital requirements were completely determined by regulatory mandate prior to the 1996 market risk amendment, this "internal models" approach for setting market risk capital requirements indicates an important change in how regulatory oversight is conducted. Having established the formula for calculating the desired capital charges, bank regulators must now evaluate the accuracy of the VaR models used to set them. In the following section, four methods for evaluating VaR model accuracy are discussed.

B. Alternative Evaluation Methods

In accordance with the current regulatory framework and for the purposes of this paper, the accuracy of VaR models is assessed with respect to their one-step-ahead forecasts; i.e., k=1. Thus, given a set of one-step-ahead VaR forecasts, regulators must determine whether the underlying model is "acceptably accurate". Three hypothesis-testing methods using different types of VaR forecasts are available; specifically, the binomial, interval forecast and distribution forecast methods. Their common premise is to determine whether the VaR forecasts in question exhibit a specified property characteristic of accurate VaR forecasts using hypothesis tests.

However, as noted by Diebold and Lopez (1996), it is unlikely that forecasts from an economic model will be fully optimal and exhibit all the properties of accurate forecasts. Thus,

[5] The 1996 market risk amendment contains a number of other qualitative criteria that banks' risk management systems must meet in order to be considered appropriate for determining market risk capital requirements.

the evaluation of a model's forecasts based on the presence of a specific statistical property will provide only limited information regarding model accuracy. In this paper, an evaluation method, based on the probability forecasting framework presented by Lopez (1997), is proposed. With this method, the relative accuracy of VaR models is evaluated by how well their probability forecasts minimize a regulatory loss function. Thus, this evaluation method can provide additional information on VaR model accuracy with respect to regulatory criteria, as opposed to the statistical criteria implied in the hypothesis-testing methods.

B.1. Evaluation of VaR estimates based on the binomial distribution

Under the current regulatory framework, banks will report their one-day, 99 percent VaR estimates (denoted $VaR_t^m(1,1)$ or simply $VaR_t^m(1)$) to their regulators, who also observe whether actual portfolio losses exceed these estimates.[6] Under the assumption that the VaR estimates are accurate, such observations can be modeled as draws from an independent binomial random variable with a probability of occurrence equal to one percent or, more generally, a specified α percent. The binomial method that regulators have chosen is based on the number of times that ε_{t+1} is less than $VaR_t^m(\alpha)$ (denoted here as x) in a sample of size T. Accurate VaR estimates should exhibit the property that their unconditional coverage, measured by $\alpha^* = x/T$, equals the desired coverage level α. Thus, the relevant null hypothesis is $\alpha^*=\alpha$, and the appropriate likelihood ratio statistic based on the binomial distribution is

$$LR_{uc}(\alpha) = 2\left[\log\left(\alpha^{*x}\left(1-\alpha^*\right)^{T-x}\right) - \log\left(\alpha^x(1-\alpha)^{T-x}\right)\right].$$

Note that the $LR_{uc}(\alpha)$ test of this null hypothesis is uniformly most powerful for a given T and that the statistic has an asymptotic $\chi^2(1)$ distribution. So, if we decide to set the size of the test at five percent, we would reject the null hypothesis if $LR_{uc}(\alpha) > 3.84$.[7]

However, the finite sample size and power characteristics of this test are of interest here.

---

[6] Note that these quantities are reported in dollar terms. Further note that VaR estimates do not capture the financial risks introduced by banks' intraday trading. Regulators are aware of such risks, but have generally chosen to monitor them using qualitative methods, such as evaluating the reasonableness of intraday position limits.

[7] Note that the size of the test, which in this paper is set at five percent, is different from the α percent coverage level of the VaR estimates in question.

With respect to size, the finite sample distribution for a specific (α,T) pair may be sufficiently different from the $\chi^2(1)$ distribution that the asymptotic critical values may be inappropriate. For this paper, the finite-sample distributions for specific (α,T) pairs are determined via simulation and compared to the asymptotic one in order to establish the actual sizes of the tests. As for power, Kupiec (1995) describes how this test generally has a limited ability to distinguish among alternative hypotheses and thus has low power, even in moderately large samples.

B.2. Evaluation of VaR interval forecasts

VaR estimates can clearly be viewed as interval forecasts of the lower left-hand tail of f_{t+1} at a specified coverage level α.[8] Interval forecasts can be evaluated conditionally or unconditionally; that is, forecast performance can be examined over the sample period with or without reference to the information available at each point in time. The $LR_{uc}(\alpha)$ test is obviously an unconditional test since it ignores this type of information. However, in the presence of the higher-moment dynamics often found in financial time series, testing for conditional accuracy is important since interval forecasts ignoring such dynamics may have correct unconditional coverage, but may have incorrect conditional coverage at any given time. As shown in Figure 1, variance dynamics can lead to clustered exceptions that may permit correct unconditional coverage but certainly not correct conditional coverage. Thus, since the $LR_{uc}(\alpha)$ test does not have power against the alternative hypothesis that the exceptions are clustered in a time-dependent fashion, it is only of limited use in the evaluation of VaR estimates.

The $LR_{cc}(\alpha)$ test proposed by Christoffersen (1998) is specifically a test of correct conditional coverage. For a given coverage level α, one-step-ahead interval forecasts are formed using model m and are denoted $V_t^m(\alpha) \equiv \left(-\infty, VaR_t^m(\alpha) \right]$. From these forecasts and the observed portfolio returns, the indicator variable $I_{t+1}^m(\alpha)$ is constructed as

$$I_{t+1}^m(\alpha) = \begin{cases} 1 \text{ if } \varepsilon_{t+1} \in V_t^m(\alpha) \\ 0 \text{ if } \varepsilon_{t+1} \notin V_t^m(\alpha) \end{cases}.$$

[8] See Chatfield (1993) for a general discussion of interval forecasts. Interval forecast evaluation techniques are also discussed by Granger, White and Kamstra (1989).

Accurate VaR interval forecasts should exhibit the property of correct conditional coverage, which implies that the $I_{t+1}^{m}(\alpha)$ series must exhibit both correct unconditional coverage and serial independence. The $LR_{cc}(\alpha)$ test of this joint hypothesis is formed by combining tests of each property. The relevant test statistic is $LR_{cc}(\alpha) = LR_{uc}(\alpha) + LR_{ind}(\alpha)$, which is distributed $\chi^2(2)$.

Note that the $LR_{ind}(\alpha)$ statistic is a likelihood ratio statistic of the null hypothesis of serial independence against the alternative of first-order Markov dependence.[9] The likelihood function under this alternative hypothesis is $L_A = \left(1-\pi_{01}\right)^{T_{00}}\pi_{01}^{T_{01}}\left(1-\pi_{11}\right)^{T_{10}}\pi_{11}^{T_{11}}$, where the T_{ij} notation denotes the number of observations in state j after having been in state i the period before, $\pi_{01} = T_{01}/\left(T_{00}+T_{01}\right)$ and $\pi_{11} = T_{11}/\left(T_{10}+T_{11}\right)$.[10] Under the null hypothesis of independence, $\pi_{01} = \pi_{11} = \pi$, and the relevant likelihood function is $L_0 = (1-\pi)^{T_{00}+T_{10}}\pi^{T_{01}+T_{11}}$, where $\pi = \left(T_{01}+T_{11}\right)/T$. Thus, the relevant test statistic is formed as $LR_{ind}(\alpha) = 2\left[\log L_A - \log L_0\right]$, which is distributed $\chi^2(1)$.

B.3. Evaluation of VaR distribution forecasts

Since VaR models are generally characterized by their forecast of f_{t+k}, Crnkovic and Drachman (1996) propose to evaluate such models based on their entire forecasted distributions. The object of interest in this evaluation method is the observed quantile q_{t+1}^{m}, which is the quantile under f_{t+1}^{m} in which the observed return ε_{t+1} actually falls; that is,

$$q_{t+1}^{m}(\varepsilon_{t+1}) = \int_{-\infty}^{\varepsilon_{t+1}} f_{t+1}^{m}(x)dx.$$

This evaluation method tests whether the observed quantiles derived under a model's distribution forecasts exhibit the properties of observed quantiles from accurate distribution forecasts. Specifically, since the quantiles of random draws from a distribution are uniformly distributed over the unit interval, the observed quantiles should be independent and uniformly distributed.

Crnkovic and Drachman (1996) suggest that these two properties be examined separately

[9] Although not done in this paper, higher-order dependence could be specified. Christoffersen (1998) also presents an alternative test of this null hypothesis based on the runs test of David (1947).

[10] Note that the formulae relating the π_{ij} variables to the transition counts are maximum likelihood estimates.

and thus propose two separate hypothesis tests.[11] As in the interval forecast method, the independence of the observed quantiles indicates whether the VaR model captures the higher-order dynamics in the return series. To test for this property, the authors suggest the use of the BDS statistic (see Brock *et al.*, 1991). However, in this paper, the focus is on their proposed test of uniform distribution.[12] The test of the uniform distribution of the q^{m}_{t+1} series is based on the Kupier statistic, which measures the deviation between two cumulative distribution functions.[13] Denoting $D_m(x)$ as the cumulative distribution function of the observed quantiles, the Kupier statistic for the deviation of $D_m(x)$ from the uniform distribution is

$$K_m = \max\left(D_m(x) - x\right) + \max\left(x - D_m(x)\right),$$

where $x \in [0,1]$. Note that for this paper, the finite sample distribution of K_m as generated in the following simulation exercise is used.[14] In general, this testing procedure is relatively data-intensive, and the authors note that test results begin to seriously deteriorate with fewer than 500 observations.

B.4. Evaluation of VaR probability forecasts

The evaluation method proposed here is based on the probability forecasting framework presented by Lopez (1997). In contrast to the hypothesis-testing methods discussed above, this method is based on standard forecast evaluation tools and gauges the accuracy of VaR models by

---

[11] Note that other authors have recently proposed other tests based on models' empirical quantiles. Diebold, Gunther and Tay (1998) propose the use of CUSUM statistics, and Berkowitz (1998) proposes likelihood ratio statistics based on a simple transformation of these quantiles.

[12] Note that the emphasis in this paper on just the second property will understate the ability of the overall evaluation method to gauge VaR model accuracy since model misclassification by the test for uniform distribution might be correctly indicated by the test for independence.

[13] Crnkovic and Drachman (1996) indicate that an advantage of the Kupier statistic is that it is equally sensitive for all values of x, as opposed to the Kolmogorov-Smirnov statistic that is most sensitive around the median. See Press *et al.* (1992) for further discussion.

[14] The asymptotic distribution of the Kupier statistic is characterized as $\text{Prob}\left(K > K_m\right) = G\left(\left[\sqrt{T} + 0.155 + 0.24/\sqrt{T}\right]v_m\right)$, where $G(\lambda) = 2\sum_{j=1}^{\infty}\left(4j^2\lambda^2 - 1\right)e^{-2j^2\lambda^2}$, $v_m = \max\left|D_m(x) - x\right|$, T is the sample size, and $x \in [0,1]$.

how well their probability forecasts minimize a regulatory loss function. Thus, by directly incorporating regulatory loss functions into the forecast evaluations, this method provides useful information on the performance of VaR models with respect to regulatory criteria as opposed to the purely statistical criteria implied by the hypothesis-testing methods.

The proposed evaluation method incorporates the interests of the regulators (or, more generally, forecast evaluators) into the forecast evaluation in two ways.[15] First, the event of interest to the regulator must be specified.[16] Thus, instead of focusing exclusively on a fixed quantile of the forecasted distributions or on the entire distributions themselves, this method allows the evaluation of VaR models based upon the regions of the distributions that are of most interest. In this paper, three types of regulatory events are considered, although many are possible.

The first type of event is whether an observed ε_{t+1} lies in the lower tail of its unconditional distribution based on past observations, denoted $\hat{F}$. Specifically, the lower α percent quantile of $\hat{F}$ is determined, and probability forecasts of whether subsequent returns will be less than it are generated. In mathematical terms, the relevant probability forecasts, conditional on the information available at time t, are

$$P_t^m = \Pr\left(\varepsilon_{t+1} < CV\left(\alpha, \hat{F}\right)\right) = \int_{-\infty}^{CV\left(\alpha, \hat{F}\right)} f_{t+1}^m(x)\,dx,$$

where $CV(\alpha, \hat{F}) = \hat{F}^{-1}(\alpha/100)$ is the unconditional quantile of interest.

The second type of event is a portfolio loss of a fixed magnitude; that is, regulators may be interested in determining how well a VaR model can forecast a portfolio loss of p percent of y_t over a one-day period. The corresponding probability forecasts generated from model m,

---

[15] Crnkovic and Drachman (1996) note that the Kupier statistic can be tailored to the interests of the forecast evaluator by introducing a user-defined weighting function.

[16] The relevance of such probability forecasts to regulators (as well as market participants) is well established. For example, Greenspan (1996b) stated that "[i]f we can obtain reasonable estimates of portfolio loss distributions, [financial] soundness can be defined, for example, as the probability of losses exceeding capital. In other words, soundness can be defined in terms of a quantifiable insolvency probability." For a more general discussion of probability forecasting in a decision theoretic framework, see Granger and Pesaran (1996).

conditional on the information available at time t, are

$$P_t^m = \Pr\left(y_{t+1} < \left(1 - \frac{p}{100} \right) y_t \right) = \Pr\left(y_t + \varepsilon_{t+1} < \left(1 - \frac{p}{100} \right) y_t \right)$$

$$= \Pr\left(\varepsilon_{t+1} < \frac{-p}{100} y_t \right) = \int_{-\infty}^{-p/100 * y_t} f_{t+1}^m(x)\,dx.$$

The third type of regulatory event corresponds to whether a bank's capital is sufficient to cover portfolio losses (in dollar terms) over a certain time period. Suppose an amount of C dollars is set aside to cover the expected maximum portfolio loss that might occur, relative to Y_τ, over the period [t+1, t+T] for $\tau \le t$. Capital C is sufficient to cover losses if $Y_i > Y_\tau - C$ $\forall i \in [t+1, t+T]$. To translate this inequality into portfolio returns, the equivalent expression $Y_i > Y_\tau e^{-\gamma(C)}$ is used, which implies that $y_i > y_\tau - \gamma(C)$. A regulator may be interested in a VaR model's ability to forecast, conditional on the information at time t, whether this capital level was not sufficient to cover portfolio losses. The corresponding probability forecast generated from model m is then

$$P_t^m = \Pr\left(y_{t+1} - y_\tau < -\gamma(C) \right) = \Pr\left(y_t + \varepsilon_{t+1} - y_\tau < -\gamma(C) \right)$$

$$= \Pr\left(\varepsilon_{t+1} < -\gamma(C) + y_\tau - y_t \right) = \int_{-\infty}^{-\gamma(C) + y_\tau - y_t} f_{t+1}^m(x)\,dx.$$

Note that this type of event does not depend on how the capital level C is determined; for example, C may be mandated by the regulators or completely determined by the bank. An interesting example of the latter case is the "precommitment" approach in which a bank reports C to the regulator and is penalized if the dollar value of portfolio losses over the following quarter at any time exceeds C; see Kupiec and O'Brien (1995) for further discussion.

The second way of incorporating regulatory interests into this evaluation method is the selection of the loss function used to evaluate the probability forecasts. Regulators should select a loss function that most directly represents their concerns. For example, the quadratic probability score (QPS), developed by Brier (1950), specifically measures the accuracy of probability forecasts over time. The QPS is the analog of mean squared error for probability

forecasts and thus implies a quadratic loss function.[17] The QPS for model m over a sample of size T is

$$QPS_m = \frac{1}{T}\sum_{t=1}^{T} 2\left(P_t^m - R_{t+1}\right)^2,$$

where R_{t+1} is an indicator variable that equals one if the specified event occurs and zero otherwise. Note that $QPS_m \in [0,2]$ and has a negative orientation such that smaller values indicate more accurate forecasts. Thus, since accurate VaR models are expected to generate lower QPS values than inaccurate models, QPS_m values closer to zero should indicate the relative accuracy of the VaR model. The QPS measure is used in this paper because it reflects the regulator's stated goal of evaluating a bank's VaR model based on the accuracy of its VaR estimates.

A key property of the QPS is that it is a strictly proper scoring rule; that is, forecasters must report their actual probability forecasts to minimize their expected QPS score. To see the importance of this property for the purpose of regulatory oversight, consider the following definition. Let P_t^m be the actual probability forecast generated by a bank's VaR model, and let $S(p_t, j)$ denote a scoring rule that assigns a numerical score to a probability forecast p_t based on whether the event occurs (j=1) or not (j=0). The reporting bank's expected score conditional on its model is

$$E\left[S\left(p_t, j\right) \mid m\right] = P_t^m S\left(p_t, 1\right) + \left(1 - P_t^m\right) S\left(p_t, 0\right).$$

The scoring rule S is strictly proper if $E\left[S\left(P_t^m, j\right) \mid m\right] < E\left[S\left(p_t, j\right) \mid m\right] \forall\ p_t \neq P_t^m$. Thus, truthful reporting is explicitly encouraged since the bank receives no benefit from modifying its actual forecasts.[18] This property is obviously important in the case of a regulator evaluating VaR models that it may not directly observe.

An important drawback of the probability forecast evaluation method is that the properties of the QPS value for a particular model and specified event cannot be easily

---

[17] Other scoring rules with different implied loss functions are available; see Murphy and Daan (1985).

[18] The scoring rule S is proper if $E\left[S\left(P_t^m, j\right) \mid m\right] \leq E\left[S\left(p_t, j\right) \mid m\right] \forall\ p_t \neq P_t^m$. Such scoring rules do not encourage the "hedging" of reported probability forecasts, but they also do not guard against it completely.

determined a priori, as opposed to the three aforementioned test statistics whose distributions are known. Thus, this evaluation method cannot be used, as the other methods, to statistically classify a VaR model as "acceptably accurate" or "inaccurate" in an absolute sense. Instead, it can be used to monitor the relative accuracy of a VaR model over time and in relation to other VaR models, which should be useful information for both model users and regulators. Although there are challenges to making this method operational, regulators may use this information on the relative accuracy of VaR models to complement that of the hypothesis-testing methods.

III. Simulation Exercise

The following simulation exercise gauges the ability of the four VaR evaluation methods to avoid model misclassification. For the three hypothesis-testing methods, this is a direct analysis of the power of these tests; i.e., determining the probability with which the tests reject the specified null hypothesis when in fact it is incorrect. If the power of a test is low, then it is very likely that the corresponding evaluation method will misclassify an inaccurate VaR model as "acceptably accurate". With respect to the probability forecast method, its ability to correctly classify VaR models is gauged by how frequently the QPS value for the true data generating process is smaller than that of the alternative models. Further analysis of the QPS values using hypothesis-testing techniques proposed by Diebold and Mariano (1995) permit a power comparison across the four evaluation methods.

The first step in this simulation exercise is determining what type of portfolio to analyze. VaR models are designed to be used with typically complicated portfolios that contain a variety of financial assets, possibly even derivatives. However, for the purposes of this exercise, the portfolio value in question is simplified to be $y_{t+1} = y_t + \varepsilon_{t+1}$, where $\varepsilon_{t+1} | \Omega_t \sim f_{t+1}$. This specification of y_{t+1} is representative of linear, deterministic conditional mean specifications. It is only for portfolios with nonlinear components, such as derivative instruments, that this choice presents inference problems; further research along these lines, as by Pritsker (1997) and Berkowitz (1998), is needed.

The simulation exercise is conducted in four distinct, yet interrelated, sections. In the first two sections, the emphasis is on the shape of the f_{t+1} distribution. To examine performance

under different distributional assumptions, the simulations are conducted by setting f_{t+1} to the standard normal distribution and a t-distribution with six degrees of freedom, which has fatter tails than the standard normal. The next two sections examine the performance of the evaluation methods in the presence of variance dynamics. Specifically, innovations from a GARCH(1,1)-normal process and a GARCH(1,1)-t(6) process are used.

In each section, the true data generating process (DGP) is one of the seven VaR models evaluated and is designated as the true model. Traditional power analysis of a hypothesis test is conducted by varying a particular parameter and determining whether the corresponding incorrect null hypothesis is rejected; such changes in parameters generate what are usually known as local alternatives. However, in this analysis, we examine alternative VaR models that are not all nested, but are commonly used in practice. Such models are here considered to be reasonable "local" alternatives, although no definitive metric is used to support this claim. For example, a popular type of VaR model specifies the variance of f_{t+1}^{m}, denoted h_{t+1}^{m}, as an exponentially weighted, moving average of squared innovations; that is,

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

This VaR model, a version of which is used in the well-known Riskmetrics calculations (see J.P. Morgan, 1995), is calibrated here by setting λ equal to 0.97 or 0.99, which imply a high-degree of persistence in variance.[19] A description of the alternative models used in each section of the simulation exercise follows.

For the first section, the true DGP is the standard normal; i.e., $\varepsilon_{t+1} | \Omega_{t} \sim N(0,1)$. The six alternative models examined are normal distributions with variances of 0.5, 0.75, 1.25 and 1.5 as well as the two calibrated VaR models with normal distributions. For the second section, the true DGP is a t(6) distribution; i.e., $\varepsilon_{t+1} | \Omega_{t} \sim t(6)$. The six alternative models are two normal distributions with variances of 1 and 1.5 (the same variance as the true DGP) and the two calibrated models with normal distributions as well as with t(6) distributions.

---

[19] Note that this model is often implemented with a finite lag-order. For example, the infinite sum is frequently truncated at 250 observations, which roughly accounts for 90 percent of the sum of the weights. See Hendricks (1996) for further discussion on the choice of λ and the truncation lag. In this paper, no such truncation is imposed, but of course, one is implied by the overall sample size of the simulated time series.

For the latter two sections, variance dynamics are introduced by using conditional heteroskedasticity of the GARCH form; i.e., $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$, which has an unconditional variance of 1.5. The only difference between the DGP's in these two sections is the chosen distributional form. For the third section, $\varepsilon_{t+1} | \Omega_t \sim N(0, h_{t+1})$, and for the fourth section, $\varepsilon_{t+1} | \Omega_t \sim t(h_{t+1}, 6)$. The six alternative models examined in these two sections are the homoskedastic models of the N(0,1), N(0,1.5) and t(6) distributions and the heteroskedastic models of the two calibrated models with normal innovations and the GARCH model with the other distributional form.

In all of the sections, the simulation runs are structured identically. The results are based on 1000 simulation runs. For each run, the simulated y_{t+1} series is generated using the chosen DGP. After 1000 initial observations and 2500 in-sample observations, the seven chosen VaR models are used to generate the specified one-step-ahead VaR forecasts for the next 500 out-of-sample observations. In the current regulatory framework, the evaluation period is set at 250 observations, but 500 observations are used here since the distribution forecast and probability forecast evaluation methods are data-intensive.

The VaR forecasts from the various models are then evaluated using the appropriate evaluation methods. For the binomial and interval forecast methods, VaR estimates for coverage levels $\alpha = [1, 5, 10]$ are examined. For the distribution forecast method, the entire forecasted distribution is examined, and for the probability forecast method, the three types of regulatory events previously discussed are examined. Specifically, for the first event, the empirical distribution function $\hat{F}$ is based on the 2500 in-sample observations, and the desired α percent critical values $CV(\alpha, \hat{F})$ are determined. The probability forecasts of whether the observed returns in the out-of-sample period will be less than $CV(\alpha, \hat{F})$ are generated for $\alpha = [1, 5, 10]$, and in the tables, these simulation results are labeled QPSe1(α). For the second event, a fixed one percent loss of log portfolio value is set as the one-day decline of interest,[20] and probability forecasts of whether the observed returns exceed that percentage loss are generated as

$$P_t^m = Pr(y_{t+1} < 0.99 y_t) = Pr(y_t + \varepsilon_{t+1} < 0.99 y_t) = Pr(\varepsilon_{t+1} < -0.01 y_t).$$

[20] Note that, in dollar terms, the event of interest is thus whether $Y_{t+1} < Y_t^{0.99}$.

In the tables, these simulation results are labeled QPSe2.

For the third event, $\gamma(C)$ is ten percent of the last in-sample log portfolio value denoted y_0; i.e., $\gamma(C) = 0.1*y_0$.[21] The choice of ten percent is related to certain regulatory reserve requirements. Thus, the probability forecast of interest is

$$P_t^m = \Pr\left(y_{t+1} - y_0 < -\gamma(C) \right) = \Pr\left(y_t + \varepsilon_{t+1} - y_0 < -0.1\, y_0 \right) = \Pr\left(\varepsilon_{t+1} < 0.9\, y_0 - y_t \right).$$

In the tables, these simulation results are labeled QPSe3. Note that given the nature of this event (i.e., whether a stochastic process ever dips below a specified barrier), it is likely the event may never occur in certain simulations. In such cases, the probability forecasts for all the models are extremely small and, to insure efficient computer simulation, are rounded down to zero whenever $P_t^m < 0.0001$. However, this adjustment can lead to QPS values exactly equal to zero, which must be accounted for in the analysis of the results. To do so, such zero-value simulation results are removed from the analysis, and the QPS analysis is based on the smaller number of simulations. The rationale behind examining these adjusted results is that model accuracy cannot be examined well if the event in question does not occur. Overall, the inference drawn from this type of regulatory event will generally be less useful due to the lower frequency of occurrence.

The main object of interest from these simulation results for the probability forecasting method is the frequency with which the QPS value for the true model is less than that for the alternative model. If this frequency is high (say, greater than 75%), then this evaluation method is generally capable of gauging model accuracy and can then be used to monitor the relative performance of VaR models over time and across models. However, since in current practice regulators have only one set of forecasts to work with, it is worthwhile to conduct a power analysis of this method to make it comparable to the other methods. To do so, the differences in QPS values, denoted $d_m = \frac{2}{T}\sum_{t=1}^{T} \left(P_t^{true} - R_{t+1} \right)^2 - \left(P_t^m - R_{t+1} \right)^2$, are examined using techniques proposed by Diebold and Mariano (1995). Specifically, the null hypothesis that $d_m \geq 0$ is tested using the statistic

$$S = \frac{d_m}{\sqrt{\hat{\sigma}_m^2 / T}} \sim N(0,1),$$

[21] Note that this choice of $\gamma(C)$ implies that $C = Y_0\left(1 - Y_0^{-0.1}\right)$.

where $\hat{\sigma}_m^2$ is an estimate of the sample variance that is robust to possible time-dependent heteroskedasticity. (Note that to test this hypothesis, the differences in QPS values must be covariance stationary, a condition determined empirically for this exercise.)

IV. Simulation Results

The simulation results are organized below with respect to the four sections of the exercise. Three general points can be made regarding the results. First, the power of the hypothesis-testing methods against the incorrect null hypotheses implied by the alternative VaR models varies considerably. In some cases, the power of the tests is high (greater than 75%), but in the majority of the cases examined, the power is poor (less than 50%) to moderate (between 50% and 75%). The results indicate that these evaluation methods are thus quite likely to misclassify inaccurate models as "acceptably accurate".

Second, the probability forecast method seems capable of gauging the accuracy of alternative VaR models relative to the true DGP. In pairwise comparisons between the true model and an alternative model, the QPS value for the true model is lower than that for the alternative model in the majority of the cases examined. However, further analysis of this method's power indicates that this performance is not superior to that of the three hypothesis-testing methods in all cases. Even though the proposed method is only as capable of differentiating between VaR models as the other methods, its ability to directly incorporate regulatory loss functions into model evaluations make it a useful complement to the statistical methods currently used in the regulatory evaluation of VaR models.

Third, for the cases in which variance dynamics are introduced, all four evaluation methods generally seem more sensitive to misspecifications of the distributional form than to misspecifications of the variance dynamics. That is, the four methods seem more capable of correctly classifying as inaccurate VaR models with correct variance dynamics and incorrect distributional shape than models with incorrect variance dynamics and correct distributional shape. Thus, these evaluation methods are likely to allow regulators to more readily detect when banks are using inappropriate distributional assumptions in their VaR models. Further

simulation work must be conducted to determine the robustness of this result.

As previously mentioned, an important issue in examining the simulation results for the statistical evaluation methods is the finite-sample size of the underlying hypothesis tests. Table 1 presents the finite-sample critical values for the three statistics examined in this paper. For the two LR tests, the quantiles corresponding to the asymptotic critical values under the finite-sample distribution are also presented. The finite-sample critical values are based on 10,000 simulations of sample size T = 500 and the corresponding α. Although discrepancies are clearly present, the differences are small. The finite-sample critical values in Table 1 are used in the power analysis that follows. The critical values for the Kupier statistic are based on 1000 simulations of sample size T = 500. Note that the critical values used with respect to the Diebold-Mariano statistics are the asymptotic normal ones since the finite-sample power properties of the normal distribution for T = 500 should be very close to the asymptotic ones.

A. Simulation results for the homoskedastic standard normal data generating process

Table 2, Panel A presents the power analysis of the three hypothesis-testing methods for a fixed test size of 5%. For the homoskedastic alternative models in the first four columns, the power results vary considerably. The power of the tests is highest for the N(0,0.5) and N(0,1.5) models that are the furthest away in variance from the true N(0,1) model. However, as this difference is diminished for the N(0,0.75) and N(0,1.25) models, the power results drop considerably, although the K test retains moderately high power. For all three tests, asymmetry arises across these alternatives; that is, the tests have relatively more power against the alternatives with lower variances than against those with higher variances. The reason for this seems to be that draws from the true DGP exceed the VaR estimates of the lower variance models more frequently and thus lead to a higher rejection rate of the false null hypothesis. With respect to the calibrated heteroskedastic models in the last two columns, the three tests have no power, due to the fact that, even though heteroskedasticity is introduced, these models and their associated empirical quantiles are quite similar to the true DGP.

Table 2, Panel B contains the five sets of comparative accuracy results for the probability forecast method. Each row presents, for each defined regulatory event, the percentage of

simulations for which the true model's QPS value is lower than that of the alternative model. In most cases, these results indicate that the QPS value for the true model is lower a high percentage of the time. Specifically, the homoskedastic alternatives are clearly found to be inaccurate with respect to the true model, and the heteroskedastic alternatives only slightly less so. Note that, as expected, the adjusted results for the third event are less sharp than for the other events, mainly due to its lower frequency of occurrence. To conduct a comparable power analysis, Panel C presents the percentage of simulations for which the null hypothesis that $d_m \geq 0$ is correctly rejected at the five percent level. Using this stricter criteria, this method's power is comparable to that of the other three methods. Overall, however, this method does seem to provide information on the relative accuracy of VaR models for this simple DGP.

B. Simulation results for the homoskedastic t(6) data generating process

Table 3, Panel A presents the power analysis of the hypothesis-testing methods. Overall, the power results are poor for the two sets of LR tests. In the majority of cases, the alternative models are incorrectly classified as "acceptably accurate" a large percentage of the time. With respect to the homoskedastic models, both LR tests generally exhibit moderate to high power against the N(0,1) model at low values of α, but poor results for the N(0,1.5) model, which has the same variance as the true DGP. The results for the K test are basically indistinguishable and moderate across these two models. With respect to the heteroskedastic models in the last four columns, the power of the LR tests against these alternatives is generally low with differences between the sets of normal and t(6) alternatives occurring at high values of α. However, the K test clearly has more power over the models based on the t(6) distribution mainly because the incorrect variance dynamics create conditional t(6) distributions much more different from the true DGP than the conditional normal distributions.

Table 3, Panel B contains the comparative accuracy results for the probability forecast method. Overall, the results indicate that a moderate to high percentage of the simulations have QPS values for the alternative models that are greater than those of the true model. With respect to the homoskedastic models, the QPS values for the N(0,1) model are more frequently higher than the true model than for the N(0,1.5) model, which has the same unconditional variance as

the true model. With respect to the heteroskedastic models, the two models based on the t(6) distribution are more clearly classified as inaccurate than the two normal models, as in Panel A. Note that, as expected, the adjusted results for the third event are less sharp than for the other events due to its lower frequency of occurrence, except for calibrated models with the t-distribution. The power results presented in Panel C mirror these results and are generally low to moderate, as are those in Panel A.

C. Simulation results for the GARCH(1,1)-normal data generating process

As presented in Table 4, Panel A, the power results of the hypothesis-testing methods seem to be closely linked to the differences between distributional assumptions. With respect to the heteroskedastic models, these tests have low power against the calibrated VaR models based on the normal distribution, since these smoothed variances are similar to the GARCH variances of the true DGP. However, the results for the GARCH-t(6) model vary greatly according to α. Both LR statistics have high power at low α, while at higher α and for the K statistical tests, the tests have low to moderate power. These results indicate that these tests have little power against alternative models characterized by close approximations of the true variance dynamics but have better power with respect to models with incorrect distributional assumptions, especially further into the tails. With respect to the homoskedastic VaR models, these methods are generally able to differentiate between the N(0,1) and t(6) models. However, the tests have little power against the N(0,1.5) model, which matches the true model's unconditional variance.

Overall, the results in Table 4, Panel B indicate that the probability forecast method is generally capable of differentiating between the true and the alternative VaR models. With respect to the homoskedastic models, the loss functions are minimized for the true model a high percentage of the time in all, but the third, regulatory events. For the heteroskedastic models, this method most clearly classifies the GARCH-t(6) model as inaccurate, even though it has the exactly correct variance dynamics. The two calibrated normal models are only moderately classified as inaccurate. Note, again, that the adjusted results for the third event are not as clear due to its less frequent occurrence. As before, the power results presented in Panel C are poor to moderate and generally in line with the other three methods, although the differences across

models are not as marked as in Panel A. These results further indicate that deviations from the true distributional form have a greater impact than misspecification of the variance dynamics.

D. Simulation results for the GARCH(1,1)-t(6) data-generating process

Table 5, Panel A presents the power analysis of the hypothesis-testing methods. The power results are again linked to the distributional assumptions used, as shown in the columns for the calibrated models. Unlike in Table 4, Panel A where their distributional assumption was correct and low power was exhibited, here the distributional assumption is incorrect and much improved power is exhibited. Thus, the misspecification of the distributional form has a significant impact on the power of these tests. However, the overall power results are still relatively poor for the three heteroskedastic models, with high power only for $LR_{uc}(1)$, where the differences in distributional form are most pronounced. The K test also has low power against these alternative models. With respect to the homoskedastic models in the first three columns, all three tests have high power; i.e., misclassification is not likely.

Table 5, Panel B again indicates that the probability forecast method is capable of differentiating between the true and the alternative models. The comparative results for the first regulatory event with $\alpha=1$ are poor, due to the fact that the empirical $CV(\alpha,\hat{F})$ values were generally so negative as to cause very few observations of the event. The results for the other events are much better. With respect to the homoskedastic alternatives in the first three columns, this method is able to correctly classify the alternative models a very high percentage of the time, indicating that incorrect variance dynamics can also be detected using this evaluation method. With respect to the three heteroskedastic alternatives, the calibrated normal models are found to generate higher QPS values a large percentage of the time, certainly higher than the GARCH-normal model that captures the dynamics correctly. Again, Panel C indicates that the power of this method against these alternative models is roughly comparable to that of the other three methods. Overall, these results indicate that although approximating or exactly capturing the variance dynamics can lead to a reduction in misclassification, distributional assumptions seem to be the dominant factor in differentiating between VaR models.

V. Conclusion

Given the increasing importance of VaR models for bank risk management and especially for regulatory capital requirements, evaluating their forecast accuracy has become a necessity. This paper examines four methods for conducting such evaluations. The evaluation methods proposed to date are based on hypothesis tests; that is, they test the null hypothesis that the VaR forecasts from a model exhibit properties characteristic of accurate VaR forecasts. If these properties are not present, then the null hypothesis of model accuracy can be rejected at the specified significance level. Although such a framework provides insight, it hinges on the tests' statistical power. As discussed by Kupiec (1995) and as shown in the simulation results above, these tests can have low power against many reasonable alternative models and thus can lead to a high degree of model misclassification. Furthermore, for the linear portfolios examined, it seems that these evaluation methods are more sensitive to misspecifications of the distributional shape than of the variance dynamics. Further research on nonlinear portfolio returns is needed.

An alternative and complementary evaluation method, based on probability forecasts, is proposed and examined here. By relying on standard forecast evaluation techniques, this evaluation method gauges the relative accuracy of VaR models by how well they minimize a loss function tailored to the user's interests; in this case, the interests of bank regulators. The simulation results indicate that this method can generally distinguish between VaR models; that is, the specified QPS score for the true model is found to be lower than that of the alternative models a high percentage of the time. However, further analysis using hypothesis-testing techniques that permit a power comparison across all four methods indicates that its power can be quite low and generally in line with that of the other three methods. Thus, even though the proposed method is only as capable of differentiating between VaR models as the other methods, its ability to directly incorporate regulatory loss functions into model evaluations make it a useful complement to the statistical methods currently used in the regulatory evaluation of VaR models.

References

Berkowitz, J., 1998. "Evaluating the Forecasts of Risk Models," Manuscript, Trading Risk Analysis Group, Federal Reserve Board of Governors.

Brier, G.W., 1950. "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 75, 1-3.

Brock, W.A., Dechert, W.D., Scheinkman, J.A. and LeBaron, B., 1991. "A Test of Independence Based on the Correlation Dimension," SSRI Working Paper #8702. Department of Economics, University of Wisconsin.

Chatfield, C., 1993. "Calculating Interval Forecasts," *Journal of Business and Economic Statistics*, 11, 121-135.

Christoffersen, P.F., 1998. "Evaluating Interval Forecasts," *International Economic Review*, 39, 841-862.

Crnkovic, C. and Drachman, J., 1996. "Quality Control," *Risk*, 9, 139-143.

David, F.N., 1947. "A Power Function for Tests of Randomness in a Sequence of Alternatives," *Biometrika*, 28, 315-332.

Diebold, F.X., Gunther, T.A. and Tay, A.S., 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863-883.

Diebold, F.X. and Mariano, R., 1995. "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-264.

Diebold, F.X. and Lopez, J.A., 1996. "Forecast Evaluation and Combination," in Maddala, G.S. and Rao, C.R., eds., *Handbook of Statistics, Volume 14: Statistical Methods in Finance*, 241-268. Amsterdam: North-Holland.

Dimson, E. and Marsh, P., 1995. "Capital Requirements for Securities Firms," *Journal of Finance*, 50, 821-851.

Estrella, A., 1995. "A Prolegomenon to Future Capital Requirements," *Federal Reserve Bank of New York Economic Policy Review*, 1, 1-12.

Federal Register, 1996. "Risk-Based Capital Standards: Market Risk," 61, 47357-47378.

Granger, C.W.J. and Pesaran, M.H., 1996. "A Decision Theoretic Approach to Forecast

Evaluation," Manuscript, Trinity College, Cambridge University.

Granger, C.W.J., White, H. and Kamstra, M., 1989. "Interval Forecasting: An Analysis Based Upon ARCH-Quantile Estimators," *Journal of Econometrics*, 40, 87-96.

Greenspan, A., 1996a. Remarks at the Financial Markets Conference of the Federal Reserve Bank of Atlanta. Coral Gables, Florida.

Greenspan, A., 1996b. Remarks at the Federation of Bankers Associations of Japan. Tokyo, Japan.

Hendricks, D., 1996. "Evaluation of Value-at-Risk Models Using Historical Data," *Federal Reserve Bank of New York Economic Policy Review*, 2, 39-69.

Hendricks, D. and Hirtle, B., 1997. "Bank Capital Requirements for Market Risk: The Internal Models Approach," *Federal Reserve Bank of New York Economic Policy Review*, December, 1-12.

J.P. Morgan, 1995. *RiskMetrics Technical Document*, Third Edition. New York: JP Morgan.

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

Kupiec, P. and O'Brien, J.M., 1995. "A Pre-Commitment Approach to Capital Requirements for Market Risk," FEDS Working Paper #95-36, Board of Governors of the Federal Reserve System.

Lopez, J.A., 1997. "Evaluating the Predictive Accuracy of Volatility Models," Research Paper #9524-R, Research and Market Analysis Group, Federal Reserve Bank of New York.

Murphy, A.H. and Daan, H., 1985. "Forecast Evaluation" in Murphy, A.H. and Katz, R.W., eds., *Probability, Statistics and Decision Making in the Atmospheric Sciences*. Boulder, Colorado: Westview Press.

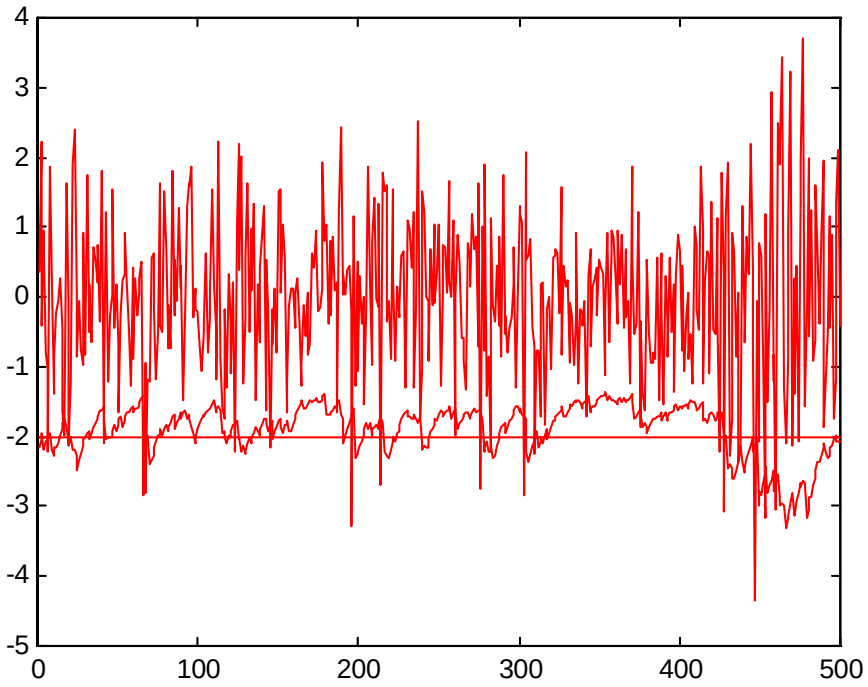
Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., 1992. *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition. Cambridge: Cambridge University Press.

Pritsker, M., 1997. "Evaluating Value at Risk Methodologies: Accuracy versus Computational Time," *Journal of Financial Services Research*, 12, 201-242.

Stahl, G., 1997. "Three Cheers," *Risk*, 10, 67-69.

Wagster, J.D., 1996. "Impact of the 1988 Basle Accord on International Banks," *Journal of Finance*, 51, 1321-1346.

Figure 1
GARCH(1,1)-Normal Process with One-Step-Ahead
Lower 5% Conditional and Unconditional Interval Forecasts



This figure graphs a realization of 500 portfolio returns from a GARCH(1,1)-normal data-generating process along with two sets of lower five percent interval forecasts. The variance dynamics are characterized as $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$, which imply an unconditional variance of 1.5. The straight line is the unconditional interval forecasts based on the unconditional N(0,1½) distribution, and the jagged line is the conditional interval forecasts based on the true data-generating process. Although both exhibit correct unconditional coverage with 25 exceptions (that is, $\alpha^* = \alpha = 5\%$), only the conditional confidence intervals exhibit correct conditional coverage or, in other words, provide 5% coverage at each point in time.

Table 1. Finite-Sample Critical Values for $LR_{uc}(\alpha)$, $LR_{cc}(\alpha)$ and K Test Statistics

	1%	5%	10%
Asymptotic $\chi^2(1)$	6.635	3.842	2.706
$LR_{uc}(1)$	7.111 (1.2%)	4.813 (7.5%)	2.613 (7.5%)
$LR_{uc}(5)$	7.299 (1.2%)	3.888 (6.3%)	3.022 (11.5%)
$LR_{uc}(10)$	7.210 (1.3%)	4.090 (6.2%)	2.887 (11.4%)
Asymptotic $\chi^2(2)$	9.210	5.992	4.605
$LR_{cc}(1)$	9.701 (1.1%)	4.801 (1.8%)	4.117 (7.0%)
$LR_{cc}(5)$	9.093 (1.0%)	5.773 (4.7%)	4.431 (9.9%)
$LR_{cc}(10)$	9.983 (1.9%)	6.237 (5.5%)	4.725 (11.3%)
K	0.0800	0.0700	0.0640

The finite-sample critical values for the $LR_{uc}(\alpha)$ and $LR_{cc}(\alpha)$ test statistics are based on 10,000 simulations of sample size T = 500. The percentages in parentheses are the quantiles that correspond to the listed asymptotic critical values under the finite-sample distributions. The finite-sample critical values for the K test statistic are based on 1,000 simulations of sample size T = 500.

Table 2.
Simulation Results for the Homoskedastic Standard Normal DGP (Units: percent)

	Model N(0,½)	N(0,¾)	N(0,1¼)	N(0,1½)	Nλ(97)	Nλ(99)
Panel A. Power of the $LR_{uc}(\alpha)$, $LR_{cc}(\alpha)$ and K Tests Against Alternative VaR Models [a]						
$LR_{uc}(1)$	99.9	54.6	32.3	70.0	3.3	6.5
$LR_{uc}(5)$	99.9	68.3	51.5	94.2	2.7	9.2
$LR_{uc}(10)$	99.9	61.5	47.4	93.1	2.3	7.3
$LR_{cc}(1)$	99.9	56.6	33.2	70.4	4.2	8.0
$LR_{cc}(5)$	99.9	64.3	40.3	89.3	3.2	9.4
$LR_{cc}(10)$	99.8	53.0	36.5	86.5	3.2	6.8
K	100	87.7	60.6	99.3	1.6	2.3
Panel B. Accuracy of VaR Models Using the Probability Forecast Method [b]						
QPSe1(1)	86.4	76.5	83.1	97.2	78.3	66.1
QPSe1(5)	98.9	84.4	82.5	97.9	80.5	74.3
QPSe1(10)	99.6	89.5	82.9	95.3	81.2	76.6
QPSe2	94.0	78.0	64.1	72.7	67.5	68.6
QPSe3[c]	57.6	48.5	66.4	73.6	60.1	60.2
Panel C. Power of the Probability Forecast Method Using the Diebold-Mariano Test [a]						
QPSe1(1)	42.8	30.3	52.3	75.0	50.6	39.7
QPSe1(5)	84.2	48.3	51.2	71.1	37.5	34.8
QPSe1(10)	89.1	55.7	34.6	64.6	45.8	35.3
QPSe2	64.6	47.3	38.6	43.9	37.1	31.3
QPSe3[c]	26.1	13.9	44.4	47.5	25.0	23.8

Notes for Table 2

Table 2 contains the simulation results for the homoskedastic standard normal DGP; that is, $\varepsilon_{t+1} | \Omega_t \sim N(0,1)$. The alternative models are normal distributions with variances of 0.5, 0.75, 1.25 and 1.5 (denoted N(0,½), N(0,¾), N(0,1¼) and N(0,1½), respectively) and normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.97$ and $\lambda = 0.99$ (denoted Nλ(97) and Nλ(99), respectively). The results are based on 1000 simulations.

Panel A presents the percentage of simulations for which the null hypothesis corresponding to each row is rejected with the test size set at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high. The $LR_{uc}(\alpha)$ rows correspond to the binomial method and examine the null hypothesis that the VaR estimates have correct unconditional coverage at the α percent level. The $LR_{cc}(\alpha)$ rows correspond to the interval forecast method and examine the null hypothesis that the VaR estimates have correct conditional coverage at the α percent level. The K row corresponds to the distribution forecast method and examines the null hypothesis that the observed quantiles are uniformly distributed.

Panel B presents the percentage of simulations for which the QPS value for the true DGP is less than that of the alternative VaR model. If this method is capable of distinguishing between the true DGP and an alternative model, then this percentage should be high. The QPSe1(α) rows correspond to the QPS values for the probability forecasts $P_t^m = \Pr\left(\varepsilon_{t+1} < CV\left(\alpha, \hat{F}\right)\right)$, where $CV\left(\alpha, \hat{F}\right)$ is the α percent quantile of the empirical cumulative distribution function $\hat{F}$. The QPSe2 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr\left(y_{t+1} < 0.99\, y_t\right)$. The QPSe3 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr\left(y_{t+1} - y_0 < -\gamma(C)\right)$, where -γ(C) is the rate of return that would reduce Y_0 to the selected capital level C.

Panel C presents the percentage of simulations for which the null hypothesis that the QPS value for the true DGP is greater than or equal to that of the alternative model is rejected at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high.

[a] The size of the tests is set at five percent.

[b] Each row represents the percentage of simulations for which the alternative model had a higher QPS value than the true DGP.

[c] The QPSe3 row has removed from it the simulations for which the QPS value of the true DGP for the third event is rounded down to zero; i.e., 23.1% of the simulations.

Table 3.
Simulation Results for the Homoskedastic t(6) DGP (Units: percent)

	Model N(0,½)	N(0,1)	Nλ(97)	Nλ(99)	tλ(97)	tλ(99)
Panel A. Power of the $LR_{uc}(\alpha)$, $LR_{cc}(\alpha)$ and K Tests Against Alternative VaR Models [a]						
$LR_{uc}(1)$	13.0	86.9	19.6	25.3	21.2	18.1
$LR_{uc}(5)$	11.5	62.1	3.8	3.1	68.1	52.7
$LR_{uc}(10)$	25.7	35.5	13.9	8.0	73.9	60.0
$LR_{cc}(1)$	14.8	89.4	20.7	15.8	26.0	33.1
$LR_{cc}(5)$	6.1	58.2	2.3	3.7	51.0	62.9
$LR_{cc}(10)$	17.3	29.9	8.7	14.0	61.2	70.9
K	69.5	49.8	57.0	64.4	97.6	98.7
Panel B. Accuracy of VaR Models Using the Probability Forecast Method [b]						
QPSe1(1)	68.1	84.9	79.1	76.6	96.3	91.0
QPSe1(5)	64.5	88.4	90.5	79.0	98.2	95.2
QPSe1(10)	76.6	79.2	90.0	80.9	97.2	94.2
QPSe2	71.7	76.2	79.7	80.4	84.0	84.1
QPSe3[c]	52.3	48.5	55.1	55.9	74.3	73.1
Panel C. Power of the Probability Forecast Method Using the Diebold-Mariano Test [a]						
QPSe1(1)	26.0	40.6	37.9	30.9	73.2	67.1
QPSe1(5)	19.6	51.4	50.0	39.0	76.1	67.6
QPSe1(10)	41.4	43.1	63.3	42.2	65.3	69.0
QPSe2	30.6	18.0	44.2	39.6	47.1	47.3
QPSe3[c]	32.4	18.2	31.6	25.7	40.4	42.6

Notes for Table 3

Table 3 contains the simulation results for the homoskedastic t(6) DGP; that is, $\varepsilon_{t+1} | \Omega_t \sim t(6)$. The alternative models are normal distributions with variances of 0.5 and 1 (denoted N(0,½) and N(0,1), respectively); normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.97$ and $\lambda = 0.99$ (denoted $N\lambda(97)$ and $N\lambda(99)$, respectively); and t(6) distributions with the same calibrated variances (denoted $t\lambda(97)$ and $t\lambda(99)$, respectively). The results are based on 1000 simulations.

Panel A presents the percentage of simulations for which the null hypothesis corresponding to each row is rejected with the test size set at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high. The $LR_{uc}(\alpha)$ rows correspond to the binomial method and examine the null hypothesis that the VaR estimates have correct unconditional coverage at the α percent level. The $LR_{cc}(\alpha)$ rows correspond to the interval forecast method and examine the null hypothesis that the VaR estimates have correct conditional coverage at the α percent level. The K row corresponds to the distribution forecast method and examines the null hypothesis that the observed quantiles are uniformly distributed.

Panel B presents the percentage of simulations for which the QPS value for the true DGP is less than that of the alternative VaR model. If this method is capable of distinguishing between the true DGP and an alternative model, then this percentage should be high. The QPSe1(α) rows correspond to the QPS values for the probability forecasts $P_t^m = \Pr\left(\varepsilon_{t+1} < CV\left(\alpha, \hat{F}\right)\right)$, where $CV\left(\alpha, \hat{F}\right)$ is the α percent quantile of the empirical cumulative distribution function $\hat{F}$. The QPSe2 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr\left(y_{t+1} < 0.99\, y_t\right)$. The QPSe3 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr\left(y_{t+1} - y_0 < -\gamma(C)\right)$, where $-\gamma(C)$ is the rate of return that would reduce Y_0 to the selected capital level C.

Panel C presents the percentage of simulations for which the null hypothesis that the QPS value for the true DGP is greater than or equal to that of the alternative model is rejected at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high.

[a] The size of the tests is set at five percent.

[b] Each row represents the percentage of simulations for which the alternative model had a higher QPS value than the true DGP.

[c] The QPSe3 row has removed from it the simulations for which the QPS value of the true DGP for the third event is rounded down to zero; i.e., 11.8% of the simulations.

Table 4.
Simulation Results for the GARCH(1,1)-Normal DGP (Units: percent)

	Model N(0,1½)	N(0,1)	t(6)	Nλ(97)	Nλ(99)	GARCH-t
Panel A. Power of the $LR_{uc}(\alpha)$, $LR_{cc}(\alpha)$ and K Tests Against Alternative VaR Models [a]						
$LR_{uc}(1)$	22.7	73.9	71.3	4.3	4.8	91.6
$LR_{uc}(5)$	30.7	73.9	72.0	5.4	6.0	81.7
$LR_{uc}(10)$	29.0	65.7	60.3	5.2	5.7	50.0
$LR_{cc}(1)$	29.3	77.2	72.8	6.1	10.9	91.6
$LR_{cc}(5)$	33.5	73.5	71.1	7.2	12.4	72.9
$LR_{cc}(10)$	29.8	63.6	60.6	6.6	11.2	39.0
K	38.6	80.6	67.6	5.5	5.4	50.5
Panel B. Accuracy of VaR Models Using the Probability Forecast Method [b]						
QPSe1(1)	60.7	66.8	79.2	50.1	51.0	93.0
QPSe1(5)	89.0	92.1	86.4	64.0	66.5	88.8
QPSe1(10)	88.9	93.3	89.9	61.6	66.1	77.1
QPSe2	82.7	85.2	85.1	60.4	63.7	64.1
QPSe3[c]	57.3	49.7	60.1	53.1	52.8	73.1
Panel C. Power of the Probability Forecast Method Using the Diebold-Mariano Test [a]						
QPSe1(1)	32.9	38.4	50.6	33.8	37.1	56.5
QPSe1(5)	60.0	64.5	56.5	46.4	52.7	60.8
QPSe1(10)	65.3	66.0	63.0	53.2	61.5	53.9
QPSe2	50.6	65.9	54.4	35.1	47.4	35.0
QPSe3[c]	24.1	22.7	28.6	23.9	27.3	55.1

Notes for Table 4

Table 4 contains the simulation results for the heteroskedastic GARCH(1,1)-normal DGP; that is, $\varepsilon_{t+1} | \Omega_t \sim N(0, h_{t+1})$, where $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$. The alternative models are normal distributions with variances of 1.5 and 1 (denoted N(0,1½) and N(0,1), respectively); a t-distribution with 6 degrees of freedom (denoted t(6)); normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.97$ and $\lambda = 0.99$ (denoted Nλ(97) and Nλ(99), respectively); and a GARCH(1,1) process with the same variance dynamics as the DGP and a t(6) distribution (denoted GARCH-t). The results are based on 1000 simulations.

Panel A presents the percentage of simulations for which the null hypothesis corresponding to each row is rejected with the test size set at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high. The $LR_{uc}(\alpha)$ rows correspond to the binomial method and examine the null hypothesis that the VaR estimates have correct unconditional coverage at the α percent level. The $LR_{cc}(\alpha)$ rows correspond to the interval forecast method and examine the null hypothesis that the VaR estimates have correct conditional coverage at the α percent level. The K row corresponds to the distribution forecast method and examines the null hypothesis that the observed quantiles are uniformly distributed.

Panel B presents the percentage of simulations for which the QPS value for the true DGP is less than that of the alternative VaR model. If this method is capable of distinguishing between the true DGP and an alternative model, then this percentage should be high. The QPSe1(α) rows correspond to the QPS values for the probability forecasts $P_t^m = \Pr(\varepsilon_{t+1} < CV(\alpha, \hat{F}))$, where $CV(\alpha, \hat{F})$ is the α percent quantile of the empirical cumulative distribution function $\hat{F}$. The QPSe2 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr(y_{t+1} < 0.99 y_t)$. The QPSe3 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr(y_{t+1} - y_0 < -\gamma(C))$, where -γ(C) is the rate of return that would reduce Y_0 to the selected capital level C.

Panel C presents the percentage of simulations for which the null hypothesis that the QPS value for the true DGP is greater than or equal to that of the alternative model is rejected at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high.

[a] The size of the tests is set at five percent.

[b] Each row represents the percentage of simulations for which the alternative model had a higher QPS value than the true DGP.

[c] The QPSe3 row has removed from it the simulations for which the QPS value of the true DGP for the third event is rounded down to zero; i.e., 19% of the simulations.

Table 5.
Simulation Results for the GARCH(1,1)-t(6) DGP (Units: percent)

	Model N(0,1½)	N(0,1)	t(6)	Nλ(97)	Nλ(99)	GARCH-N
Panel A. Power of the $LR_{uc}(\alpha)$, $LR_{cc}(\alpha)$ and K Tests Against Alternative VaR Models [a]						
$LR_{uc}(1)$	60.8	100.0	96.4	85.8	87.1	86.5
$LR_{uc}(5)$	75.5	100.0	96.9	60.3	63.2	62.1
$LR_{uc}(10)$	80.4	100.0	96.0	36.8	38.5	39.3
$LR_{cc}(1)$	87.5	99.8	96.8	35.1	46.1	87.6
$LR_{cc}(5)$	99.5	100.0	96.9	12.8	36.7	58.4
$LR_{cc}(10)$	98.9	100.0	95.9	27.4	56.0	27.4
K	98.7	100.0	98.2	45.4	49.6	50.6
Panel B. Accuracy of VaR Models Using the Probability Forecast Method [b]						
QPSe1(1)	60.7	49.3	49.3	46.3	46.7	41.7
QPSe1(5)	99.6	91.8	90.8	84.2	84.0	69.9
QPSe1(10)	100.0	98.6	98.2	90.4	90.6	76.4
QPSe2	93.2	96.2	95.6	82.8	83.0	69.9
QPSe3[c]	63.0	66.5	64.1	55.6	55.5	45.7
Panel C. Power of the Probability Forecast Method Using the Diebold-Mariano Test [a]						
QPSe1(1)	34.9	34.9	33.9	37.5	34.9	62.7
QPSe1(5)	81.1	81.1	52.8	76.6	80.0	26.7
QPSe1(10)	95.0	95.0	62.5	85.7	94.1	5.3
QPSe2	88.8	91.9	47.3	81.7	90.7	9.7
QPSe3[c]	47.5	43.1	50.0	59.3	48.5	42.6

Notes for Table 5

Table 5 contains the simulation results for the heteroskedastic GARCH(1,1)-t(6) DGP; that is, $\varepsilon_{t+1} | \Omega_t \sim t\left(h_{t+1}, 6\right)$, where $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$. The alternative models are normal distributions with variances of 1.5 and 1 (denoted N(0,1½) and N(0,1), respectively); a t-distribution with 6 degrees of freedom (denoted t(6)); normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.97$ and $\lambda = 0.99$ (denoted Nλ(97) and Nλ(99), respectively); and a GARCH(1,1) process with the same variance dynamics as the DGP and a normal distribution (denoted GARCH-N). The results are based on 1000 simulations.

Panel A presents the percentage of simulations for which the null hypothesis corresponding to each row is rejected with the test size set at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high. The $LR_{uc}(\alpha)$ rows correspond to the binomial method and examine the null hypothesis that the VaR estimates have correct unconditional coverage at the α percent level. The $LR_{cc}(\alpha)$ rows correspond to the interval forecast method and examine the null hypothesis that the VaR estimates have correct conditional coverage at the α percent level. The K row corresponds to the distribution forecast method and examines the null hypothesis that the observed quantiles are uniformly distributed.

Panel B presents the percentage of simulations for which the QPS value for the true DGP is less than that of the alternative VaR model. If this method is capable of distinguishing between the true DGP and an alternative model, then this percentage should be high. The QPSe1(α) rows correspond to the QPS values for the probability forecasts $P_t^m = \Pr\left(\varepsilon_{t+1} < CV\left(\alpha, \hat{F}\right)\right)$, where $CV\left(\alpha, \hat{F}\right)$ is the α percent quantile of the empirical cumulative distribution function $\hat{F}$. The QPSe2 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr\left(y_{t+1} < 0.99 y_t\right)$. The QPSe3 row corresponds to the QPS values for the probability forecasts $P_t^m = \Pr\left(y_{t+1} - y_0 < -\gamma(C)\right)$, where $-\gamma(C)$ is the rate of return that would reduce Y_0 to the selected capital level C.

Panel C presents the percentage of simulations for which the null hypothesis that the QPS value for the true DGP is greater than or equal to that of the alternative model is rejected at the five percent level. If a test exhibits power against an alternative model, then this percentage should be high.

[a] The size of the tests is set at five percent.

[b] Each row represents the percentage of simulations for which the alternative model had a higher QPS value than the true DGP.

[c] The QPSe3 row has removed from it the simulations for which the QPS value of the true DGP for the third event is rounded down to zero; i.e., 6% of the simulations.