Regulatory Evaluation of Value-at-Risk Models

Jose A. Lopez

Research and Market Analysis Group
Federal Reserve Bank of New York
33 Liberty Street
New York, NY 10045
(212) 720-6633
jose.lopez@ny.frb.org

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; i.e., models of the time-varying distributions of portfolio returns. Currently, regulators have available three hypothesis-testing methods for evaluating the accuracy of VaR models: the binomial method, the interval forecast method and the distribution forecast method. These methods use hypothesis tests to examine whether the VaR forecasts in question exhibit properties characteristic of accurate VaR forecasts. However, given the low power often exhibited by these tests, these methods may often misclassify forecasts from inaccurate models as accurate. A new evaluation method that uses loss functions based on probability forecasts, is proposed. Simulation results indicate that this method is capable of differentiating between forecasts from accurate and inaccurate, alternative 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 New York or the Federal Reserve System. I thank Beverly Hirtle, Peter Christoffersen, Frank Diebold, Darryl Hendricks, Paul Kupiec, Jim O'Brien and Philip Strahan as well as participants at the 1996 Federal Reserve System Conference 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.

> 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 such activity, 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 proposes 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 have adopted this amendment and will begin enforcing it in 1998.[1] Under the amended capital rules, market risk capital charges can be based on the "value-at-risk" (VaR) estimates generated by banks' own VaR models. In general, VaR models are models of the time-varying distributions of portfolio returns, and VaR estimates are forecasts of the maximum portfolio value that could be lost over a given holding period with a specified confidence level; i.e., a specified lower quantile of the 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 in the literature. In each of these tests, the null hypothesis is that the VaR forecasts in question exhibit a specified property characteristic of

[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).

accurate VaR 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 exhibit correct unconditional coverage. The interval forecast method proposed by Christoffersen (1997) examines whether the VaR estimates exhibit correct conditional coverage, and the distribution forecast method proposed by Crnkovic and Drachman (1996) examines whether observed empirical quantiles derived from a VaR model's distribution forecast are independent and uniformly distributed. In these tests, if the null hypothesis is rejected, the VaR forecasts do not display 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 model as acceptably accurate will be high. This paper examines this issue within the context of a Monte Carlo simulation exercise using several data generating processes.

In addition, this paper proposes an alternative evaluation method based on the probability forecasting framework presented by Lopez (1997). In contrast to those listed above, this method is not based on a hypothesis testing framework, but instead attempts to determine the accuracy of VaR models using standard forecast evaluation techniques for probability forecasts. That is, the accuracy of a VaR model is gauged by how well probability forecasts from the model minimize a loss function that represents the user's interests. In this paper, the VaR probability forecasts used are of a specified regulatory event, and the loss function used is the quadratic probability score, a proper scoring rule.[2] Although the issue of statistical power is not relevant for this method, the issues of model misclassification and the comparative accuracy of VaR models under a specified loss function are examined within the context of a simulation exercise.

The simulation results indicate that the three hypothesis-testing methods can have relatively low power and thus a relatively high chance of misclassifying an inaccurate VaR model

[2] Note that a proper scoring rule, which will be defined more precisely later, is simply a loss function for which a forecaster must report their actual probability forecasts to minimize their expected score.

as "acceptably accurate". With respect to the probability forecast method, the simulation results indicate that such a technique is capable of distinguishing between the forecasts generated by the true data-generating process (and thus the accurate VaR model) and alternative models. This ability, as well as its flexibility with respect to the specification of the regulatory loss function, make a reasonable case for the use of probability forecast evaluation techniques 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 as well as the four evaluation methods examined. Sections III and IV outline the simulation experiment 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, and $y_t = \ln(Y_t)$. 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}; that is, $\varepsilon_{t+k} \mid \Omega_t \sim f_{t+k}$. Thus, VaR model m is characterized by f_{mt+k}, its forecast of f_{t+k}.

Currently, the VaR estimate is the most common type of forecast generated from VaR models. A VaR estimate is a specified quantile of a portfolio's forecasted return distribution over a given holding period. The VaR estimate at time t derived from model m for a k-period-ahead return is denoted $VaR_{mt}(k,\alpha)$ and is the critical value of f_{mt+k} that corresponds to its lower α percent tail. Thus, $VaR_{mt}(k,\alpha)$ is the solution to

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

or, equivalently, $VaR_{mt}(k,\alpha) = F^{-1}_{mt+k}(\alpha/100)$, where F_{mt+k} is the forecasted cumulative distribution function.

Given their role in internal, bank risk management and now in regulatory capital calculations, the evaluation of VaR estimates and the models generating 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.[3] 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. 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. Hence, although individual banks and regulators may use similar evaluation methods, the regulatory evaluation of VaR models has unique characteristics that need to be addressed.

In this section, the current regulatory framework for market risk capital charges is described, and several methods for the regulatory evaluation of VaR models are introduced. 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 for probability forecasts; that is, the accuracy of a VaR model is gauged by how well the specified regulatory loss function is minimized by the model's probability forecasts.

A. Current Regulatory Framework

The current U.S. risk-based capital standards for the market risk exposure of commercial banks are based on an amendment to the 1988 Basle Capital Accord. The capital standards 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. Beginning in 1998, any bank or bank holding company whose trading activity equals greater than 10 percent of its total assets or whose trading activity equals greater than $1 billion must hold regulatory capital against their market risk exposure.

Under the so-called "internal models" approach, these capital charges are based on the VaR estimates generated by banks' own internal, risk measurement models using the

---

[3] 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).

standardizing regulatory parameters of a ten-day holding period (k = 10) and 99 percent coverage (α = 1). Thus, a bank's market risk capital charge is based on its estimate of the potential loss that would not be exceeded with one percent certainty over the subsequent two week period. Specifically, a bank's market risk capital charge for time t+1, MRC_{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,

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

where S_{mt} and SR_{mt} are a regulatory multiplication factor and an additional capital charge for the portfolio's idiosyncratic credit risk, respectively.[4] Note that, under the current regulatory framework, $S_{mt} \geq 3$.

The S_{mt} multiplier is included in the calculation of market risk capital charges 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. Second, S_{mt} explicitly links the accuracy of a bank's VaR model to its capital charge by varying over time. In the current regulatory framework, S_{mt} 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 $VaR_{mt}(1,1)$.

The value of S_{mt} depends on the number of exceptions (defined as the occasions when $\varepsilon_{t+k} < VaR_{mt}(1,1)$) observed over the last 250 trading days. To address the low power of the implied, binomial hypothesis test, the number of such exceptions is divided into three zones. Within the green zone (four or fewer exceptions), a VaR model is deemed "acceptably accurate", and S_{mt} remains at three, the level specified by the Basle Committee. Within the yellow zone (five through nine exceptions), S_{mt} increases incrementally with the number of exceptions. Within the red zone (ten or more exceptions), the VaR model is deemed to be inaccurate, and S_{mt} increases to four. The institution must also explicitly improve its risk management system.

Clearly, the "internal models" approach for setting market risk capital requirements

[4] The specific risk capital charge is set in place to cover adverse price changes due to unanticipated events, such as an unexpected bond default. Although an important topic, specific risk is not the subject of this paper.

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. Below, 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 will be 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, the evaluation of a model's forecasts based on the presence of a specific property may provide only limited information regarding model accuracy. In addition, the power of the hypothesis test used in the evaluation is also an important issue to consider. In this paper, an alternative evaluation method, based on the probability forecasting framework presented by Lopez (1997), is proposed. With this method, the accuracy of VaR models is evaluated by how well probability forecasts generated by the models minimize a regulatory loss function. Thus, this evaluation method provides information on model accuracy that is directly tailored to the interests of the regulators.

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

Under the current regulatory framework, banks will report their one-day VaR estimates (i.e., $VaR_{mt}(1,\alpha) \equiv VaR_{mt}(\alpha)$) to the regulators, who also observe whether actual portfolio losses exceed these estimates. 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 the specified α percent.

As discussed by Kupiec (1995), a variety of tests are available to examine whether the observed probability of occurrence, also known as unconditional coverage, equals α, and the method that regulators have chosen is based on the number of occasions where $\varepsilon_{t+1} < VaR_{mt}(\alpha)$ in a sample. The probability of observing x such exceptions in a sample of size T is

$$\Pr(x; \alpha, T) = \binom{T}{x} \alpha^x (1-\alpha)^{T-x}.$$

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 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.

However, the finite sample size and power characteristics of this test are of interest here. 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. The finite-sample distribution for a specific (α,T) pair can be determined via simulation and compared to the asymptotic one in order to establish the actual size of the test. As for power, Kupiec (1995) describes how this test has a limited ability to distinguish among alternative hypotheses and thus low power, even in moderately large samples. Specifically, for sample sizes of regulatory interest (approximately 250 to 500 trading days) and small values of α, the power of this test in cases where α^* is just 90 percent of the true α generally does not exceed 10 percent.

B.2. Evaluation of VaR interval forecasts

VaR estimates can clearly be viewed as interval forecasts; that is, forecasts of the lower left-hand interval of f_{t+k}, the k-step-ahead return distribution, at a specified coverage level α.[5]

[5] Interval forecast evaluation techniques are also proposed by Granger, White and Kamstra (1989); see Chatfield (1993) for a general discussion of interval forecasts.

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 of interval forecasts since it ignores this type of information. However, in the presence of the time-dependent heteroskedasticity often found in financial time series, testing the conditional accuracy of interval forecasts becomes important. The main reason being that interval forecasts that ignore such variance dynamics might have correct unconditional coverage (i.e., $\alpha^* = \alpha$), but at any given time, may have incorrect conditional coverage; see Figure 1 for an illustration. Thus, the $LR_{uc}(\alpha)$ test does not have power against the alternative hypothesis that the exceptions are clustered in a time-dependent fashion.

The $LR_{cc}(\alpha)$ test proposed by Christoffersen (1997) is 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 as $V_{mt}(\alpha) \equiv \left[-\infty, VaR_{mt}(\alpha) \right]$. From these forecasts and the observed portfolio returns, the indicator variable $I_{mt}(\alpha)$ is constructed as

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

Accurate VaR interval forecasts should exhibit the property of correct conditional coverage, which implies that the $\left\{ I_{mt}(\alpha) \right\}_{t=1}^{T}$ series must exhibit both correct unconditional coverage and serial independence. The $LR_{cc}(\alpha)$ test for correct conditional coverage is formed by combining tests of correct unconditional coverage and independence, and 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.[6] 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 $\pi_{11} = T_{11} / \left(T_{10}+T_{11}\right)$. Under the null hypothesis of independence,

[6] Although not done in this paper, higher-order dependence could be specified. Christoffersen (1997) also presents an alternative test of this null hypothesis based on the runs test of David (1947).

$\pi_{01} = \pi_{11} = \pi$, 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 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}, the distribution of k-step-ahead portfolio returns, Crnkovic and Drachman (1996) propose to evaluate such models based on their entire forecasted distributions; see Diebold *et al.* (1997) for further discussion. The object of interest in this evaluation method is the observed quantile, which is the quantile under f_{mt+1} in which the observed return ε_{t+1} actually falls; i.e., given f_{mt+1} and the observed ε_{t+1}, the corresponding observed quantile is

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

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 and thus propose two separate hypothesis tests. 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.[7] The test of the uniform distribution of the $\left\{q_{mt}\right\}_{t=1}^{T}$ series is based on the Kupier statistic, which measures the deviation between two cumulative distribution functions.[8]

---

[7] Note that the emphasis on the second property will understate the power of the overall evaluation method since model misclassification by the test for uniform distribution might be correctly indicated by the test for independence.

[8] 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*

Let $D_m(x)$ denote the cumulative distribution function of the observed quantiles. The Kupier statistic for the deviation of $D_m(x)$ from the uniform distribution is

$$K_m = \max_{0 \le x \le 1}\left(D_m(x) - x\right) + \max_{0 \le x \le 1}\left(x - D_m(x)\right),$$

and its asymptotic distribution is characterized as

$$\text{Prob}\left(K > K_m\right) = G\left(\left[\sqrt{T} + 0.155 + \frac{0.24}{\sqrt{T}}\right]v_m\right),$$

where $G(\lambda) = 2\sum_{j=1}^{\infty}\left(4j^2\lambda^2 - 1\right)e^{-2j^2\lambda^2}$, $v_m = \max_{0 \le x \le 1}\left|D_m(x) - x\right|$, and T is the sample size. Note that for this paper, the finite sample distribution of K_m as generated in the following simulation exercise is used. 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). As opposed to the hypothesis-testing methods discussed previously, this method is based on standard forecast evaluation tools. That is, the accuracy of VaR models is gauged by how well their probability forecasts of specified regulatory events minimize a loss function relevant to regulators. Although statistical power is not relevant within this framework, the degree of model misclassification exhibited by this evaluation method can be examined within the context of a Monte Carlo simulation exercise.

The proposed evaluation method can be tailored to the interests of the regulators (or, more generally, forecast evaluators) in two ways.[9] First, the event of interest to the regulator must be specified.[10] Thus, instead of focussing exclusively on a fixed quantile of the forecasted

---

al. (1992) for further discussion.

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

[10] 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

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.

The first type of event is similar to the one examined above; that is, whether an observed ε_{t+1} lies in the lower tail of its forecasted distribution. Using the unconditional distribution of ε_{t+1} based on past observations (denoted F), the portfolio loss associated with the desired empirical quantile 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

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

where $CV(\alpha, F) = F^{-1}(\alpha/100)$.

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, conditional on the information available at time t, are

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

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

The third type of regulatory event corresponds to whether a bank's capital is sufficient to cover portfolio losses over a certain time period. Specifically, suppose an amount of capital, denoted C, is set aside to cover the expected maximum portfolio loss relative to Y_τ that might occur over the period [t+1, t+T] with $\tau \le t$; that is, $Y_i - Y_\tau > -C \quad \forall i \in [t+1, t+T]$. To translate this inequality into log returns, the expression $Y_i > Y_\tau - C$ is converted to the

probability forecasting in a decision theoretic framework, see Granger and Pesaran (1996).

equivalent expression $Y_i > Y_\tau e^{-\gamma(C)}$, 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 the capital level is sufficient or, in other words, whether $y_{t+1} - y_\tau$ will be less than $-\gamma(C)$. The corresponding probability forecast generated from model m is

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

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

Note that this type of event is independent of 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 exceed C; see Kupiec and O'Brien (1995) for further discussion.

The second way of tailoring the probability forecast evaluation to the interests of the regulators is the selection of the loss function or scoring rule used to evaluate the forecasts. Scoring rules measure the "goodness" of the forecasted probabilities, as defined by the forecast user. Thus, a regulator's economic loss function should be used to select the scoring rule with which to evaluate the probability forecasts. 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.[11] The QPS for model m over a sample of size T is

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

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 (i.e., smaller values indicate more accurate forecasts). Thus, accurate VaR models are expected to generate lower QPS scores than inaccurate models.

[11] Other scoring rules, such as the logarithmic score, with different implied loss functions are available; see Murphy and Daan (1985) for further discussion.

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; see also Murphy and Daan (1985). Let P_{mt} 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

$$E\left[S\left(p_t, j \right) \mid m \right] = P_{mt} S\left(p_t, 1 \right) + \left(1 - P_{mt} \right) S\left(p_t, 0 \right).$$

The scoring rule S is strictly proper if $E\left[S\left(P_{mt}, j \right) \mid m \right] < E\left[S\left(p_t, j \right) \mid m \right] \; \forall \; p_t \neq P_{mt}$. Thus, truthful reporting is explicitly encouraged since the bank receives no benefit from modifying their actual forecasts.[12] This property is obviously important in the case of a regulator evaluating VaR models that it may not directly observe.

The QPS measure is specifically used in this paper because it reflects the regulators' loss function with respect to VaR model evaluation. As outlined in the 1996 market risk amendment, the regulator's goal in collecting one-day VaR estimates is to evaluate the quality and accuracy of a bank's VaR model. Since the accuracy of such a model is an input into the regulatory capital requirement MRC_t, the regulators should specify a loss function, such as QPS, that explicitly measures forecast accuracy.

It should be noted that a drawback of the probability forecast evaluation method is that the properties of the QPS value for a particular model for a specified event cannot be 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 directly classify a VaR model as "acceptably accurate" or "inaccurate" in an absolute sense. Instead, this method can only be used to monitor the accuracy of a VaR model over time and in relation to other VaR models. Yet, as will be shown in the simulation results below, this method is well capable of gauging model accuracy, unlike the statistical tests that can have low power against reasonable alternatives. A regulator (or any other evaluator of VaR models) thus may use this proposed

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

method to provide reliable information on the relative accuracy of a VaR model.

III. Simulation Exercise

The following simulation exercise analyzes the ability of the four VaR evaluation methods to gauge the accuracy of alternative VaR models (i.e., models other than the true data generating process) and thus avoid model misclassification. For the three hypothesis-testing methods, this amounts to analyzing the power of the 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, alternative 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 lower than that of the alternative models.

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 of financial assets that can include currencies, interest-sensitive instruments and financial 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 is needed.

The simulation exercise is conducted in four distinct, yet interrelated, segments. In the first two segments, 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 induces fatter tails than the normal. The second two segments 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 segment, the true data generating process (DGP) is one of the seven VaR models evaluated and is designated as the "true" model or model 1. 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 termed local alternatives. However, in this analysis, we examine alternative VaR models that are not all nested, but are commonly used in practice and hence are reasonable "local" alternatives. For example, a popular type of VaR model specifies the variance h_{mt+1} of f_{mt+1} as an exponentially weighted, moving average of squared innovations; that is,

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

This VaR model, a version of which is used in the well-known Riskmetrics calculations (see J.P. Morgan, 1995), is calibrated in this paper by setting λ equal to 0.97 or 0.99, which imply a high-degree of persistence in variance.[13] A description of the alternative models used in each segment of the simulation exercise follows.

For the first segment, the true DGP for f_{t+1} 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 segment, the true DGP for f_{t+1} 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.

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

---

[13] Note that this model is often implemented with a finite lag-order. For example, the infinite sum is frequently truncated at 250 observations, which accounts for over 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.

In all of the segments, the simulation runs are structured identically. 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 results are based on 1000 simulation runs.

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 α = [1, 5, 10, 25] 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 F is based on the 2500 in-sample observations, and the desired α percent critical values CV(α,F) are determined. The probability forecasts of whether the observed returns in the out-of-sample period will be less than CV(α,F) are generated as

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

The four empirical quantiles examined are α =[1, 5, 10, 25], 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, and probability forecasts of whether the observed returns exceed that percentage loss are generated.[14] Thus,

$$P_{mt} = \Pr\left(y_{t+1} < 0.99\,y_t\right) = \Pr\left(y_t + \varepsilon_{t+1} < 0.99\,y_t\right) = \Pr\left(\varepsilon_{t+1} < -0.01\,y_t\right).$$

In the tables, these simulation results are labeled QPSe2.

For the third event, γ(C) is set to be ten percent of the last in-sample log portfolio value

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

denoted y_0; i.e., $\gamma(C) = 0.1*y_0$.[15] The choice of ten percent is related to actual regulatory reserve requirements. Thus, the probability forecast of interest is

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

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 that, in certain simulations, the event may never occur. In such cases, the individual probability forecasts for all the models may be extremely small and, to insure efficient computer simulation, are rounded down to zero whenever $P_{mt} < 0.0001$. However, this adjustment obviously can lead to QPS values exactly equal to zero, a result that 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 results are reported with respect to 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.

IV. Simulation Results

The simulation results are organized below with respect to the four segments of the simulation exercise. Three general points can be made regarding the results. First, the power of the hypothesis-testing methods against the incorrect null hypothesis 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 well capable of gauging the accuracy of VaR models relative to the true DGP. That is, in pairwise comparisons between the "true" model

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

and an alternative model, the QPS for the "true" model is lower than that for the alternative model in the majority of the cases examined. Thus, the chances of model misclassification when using this evaluation method would seem to be low. Given this ability to gauge relative model accuracy as well as the flexibility introduced by the specification of the regulatory loss function, a reasonable case can be made for the use of probability forecast evaluation techniques in the regulatory evaluation of VaR models.

Third, for the cases in which variance dynamics are introduced (i.e., the third and fourth sets of results), 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 differentiating between alternative models with the correct variance dynamics and different distributional shape than between alternative models with incorrect variance dynamics and the correct distributional shape. This result seems to indicate that these evaluation methods are more likely to allow regulators 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 corresponding critical values from their asymptotic distributions are also presented. These 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.

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 (models 2 through 5), the power results vary considerably. The power of the tests is high for the N(0,0.5) and N(0,1.5) models (models 2 and 5) that are the furthest away in variance from the true N(0,1) model. However, as

this difference is diminished in models 3 and 4, 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 (models 2 and 3) than against those with higher variances (models 4 and 5). 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 (models 6 and 7), the three tests have no power, due to the fact that, even though heteroskedasticity is introduced, the models and their associated empirical quantiles are quite similar to the true DGP.

Table 2, Panel B contains the six sets of comparative accuracy results for the probability forecast method. The table presents, for each defined regulatory event, the frequency with which the "true" model's QPS score is lower than the alternative model's score. Clearly, in most cases, this method indicates that the QPS score for the "true" model is lower a high percentage of the time (over 75%). 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. Overall, this evaluation method is capable of avoiding the misclassification of inaccurate 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 low for the two LR tests; that is, in the majority of cases, the chosen alternative models are incorrectly classified as "acceptably accurate" a large percentage of the time. However, the power of the K test is significantly higher against these models, mainly due to the important differences in the shapes of the alternative models' f_{mt+1} forecasts with respect to the true t(6) distribution.

With respect to the homoskedastic models, both LR tests generally exhibit moderate to high power against the N(0,1) model (model 3) at low values of α, but poor results for the N(0,1.5) model (model 2), which has the same variance as the true DGP. The results for the K

test are basically indistinguishable across these two models. With respect to the heteroskedastic models (models 4 through 7), the power of the LR tests against these alternatives is generally low with only slight differences between the sets of normal and t(6) alternatives. However, the K test clearly has more power over the models based on the t(6) distribution (models 6 and 7) 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 this method correctly gauges the accuracy of the alternative models examined; that is, a moderate to high percentage of the simulations indicate that the loss incurred by the incorrect, alternative models is greater than that of the "true" model. With respect to the homoskedastic models, this method classifies the N(0,1) model (model 3) as inaccurate more frequently than the N(0,1.5) model (model 2), 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 (models 6 and 7) are more clearly classified as inaccurate than the two normal models (models 4 and 5), 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, again except for models 6 and 7.

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

Table 4, Panel A presents the power analysis of the hypothesis-testing methods. The power results seem to be closely linked to the differences between the distributional assumptions used. Specifically, with respect to the heteroskedastic models, these tests have low power against the calibrated VaR models based on the normal distribution (models 5 and 6), mainly due to the fact that these smoothed variances are similar to the GARCH(1,1) variances of the true DGP. However, the results for the GARCH-t(6) alternative model (model 7) vary greatly according to α; that is, 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.

With respect to the homoskedastic VaR models, these methods are generally able to differentiate between the N(0,1) and t(6) models (models 3 and 4). However, the tests have little power against the N(0,1.5) model (model 2), which matches the "true" model's unconditional variance.

Table 4, Panel B presents the comparative accuracy results for the probability forecast method. Overall, the results indicate that this method is generally capable of differentiating between the "true" model and the alternative models. With respect to the homoskedastic models (models 2 through 4), the loss functions are minimized for the "true" model a high percentage of the time in all, but the third, regulatory events. With respect to the heteroskedastic models, the method most clearly classifies the GARCH-t(6) model (model 7) as inaccurate, even though it has the exactly correct variance dynamics. The two calibrated normal models (models 5 and 6) are only moderately classified as inaccurate. These results indicate that deviations from the true distributional form have a greater impact than misspecification of the variance dynamics, especially in the tail (i.e., low α events of interest). Note, again, that the adjusted results for the third event are not as clear due to its less frequent occurrence.

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. The key to seeing this result is contained in the columns for the calibrated models (models 5 and 6). Unlike in Panel A of Table 4 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, under the same misspecification of the variance dynamics, has a significant impact on the power of these tests.

However, the overall power results are still relatively poor for the heteroskedastic models (models 5 thorough 7), with high power only under the null hypothesis for α=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 (models 2 through 4), all three tests have high power; i.e., misclassification is not likely.

Table 5, Panel B presents the comparative accuracy results for the probability forecast

method. Again, the results indicate that this method is capable of differentiating between the "true" model and the alternative models. The comparative results for the first regulatory event with $\alpha=1$ are poor, but those for the other events are much better, due to the fact that the empirical CV(1,F) values were generally so negative as to cause very few observations of the event. With respect to the homoskedastic alternatives (models 2 through 4), this method is able to accurately classify the alternative models a very high percentage of the time; thus, indicating that incorrect variance dynamics can also be detected using this evaluation method. With respect to the heteroskedastic alternatives (models 5 through 7), the calibrated normal models (models 5 and 6) are found to generate higher scores a high percentage of the time, certainly higher than the GARCH-normal model (model 7) that captures the dynamics correctly. These results indicate that although approximating or exactly capturing the variance dynamics can lead to a reduction in misclassification, distributional assumptions seems to be the dominant factor in differentiating between models.

V. Conclusion

This paper addresses the question of how the users of VaR forecasts, such as banks and their regulators, can evaluate the accuracy of the underlying VaR models. The evaluation methods proposed to date are based on hypothesis tests; that is, if the VaR model is accurate, its VaR forecasts should exhibit properties characteristic of accurate VaR forecasts. If these properties are not present, then we can reject the implied null hypothesis of model accuracy at the specified significance level. Although such a framework may provide insight, it hinges on the tests' statistical power. As discussed by Kupiec (1995) and as shown in the results contained in this paper, these tests can have low power against many reasonable alternative models and thus could lead to a high degree of model misclassification.

An alternative evaluation method, based on probability forecasts, is proposed and examined. By relying on standard forecast evaluation tools, this method gauges the accuracy of VaR models by how well they minimize a loss function tailored to the user's interests; in case, the interests of bank regulators. The simulation results indicate that this method can distinguish between VaR models; that is, the probability forecast method seems to be less prone to model

misclassification. In general, it seems more sensitive to misspecifications of the distributional shape than of the variance dynamics. Given this ability to gauge model accuracy as well as the flexibility introduced by the specification of regulatory loss functions, a reasonable case can be made for the use of probability forecast evaluation techniques in the regulatory evaluation of VaR models.

References

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., 1997. "Evaluating Interval Forecasts," Manuscript, Research Department, International Monetary Fund.

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., 1997. "Evaluating Density Forecasts," Manuscript, Department of Economics, University of Pennsylvania.

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.

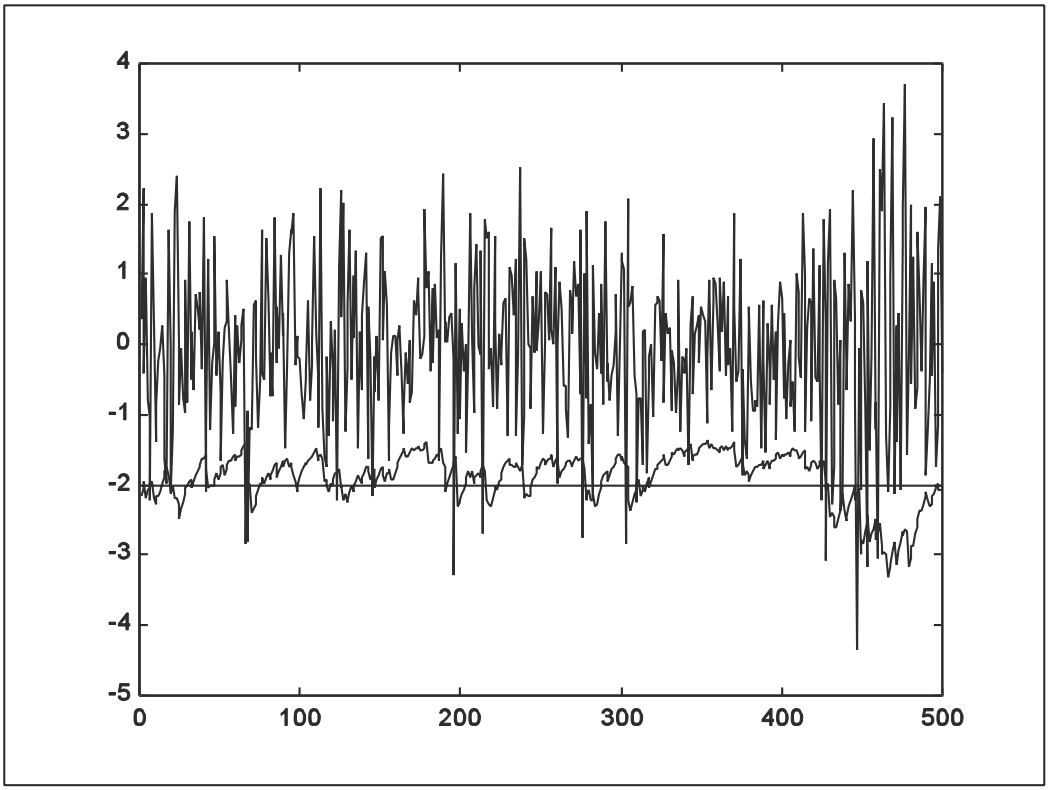
Lopez, J.A., 1997. "Evaluating the Predictive Accuracy of Volatility Models," Research Paper #9524-R, Research and Market Analysis Group, Federal Reserve Bank of New York.

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.

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 Confidence Intervals



This figure graphs a realization of length 500 of a GARCH(1,1)-normal process along with two sets of lower 5% confidence intervals. The variance dynamics are characterized as $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$, which imply an unconditional variance of 1.5. The straight line is the unconditional confidence interval based on the unconditional distribution, and the jagged line is the conditional confidence intervals based on the true data-generating process. Although both exhibit correct unconditional coverage (i.e., $\alpha^*=\alpha=5\%$), only the conditional confidence intervals exhibit correct conditional coverage.

Table 1. Finite-Sample Critical Values of $LR_{uc}(\alpha)$, $LR_{cc}(\alpha)$ 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%)
$LR_{uc}(25)$	6.914 (1.1%)	3.993 (5.1%)	2.815 (10.2%)
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%)
$LR_{cc}(25)$	9.541 (1.2%)	6.223 (5.6%)	4.727 (10.6%)
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 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 Homoskedastic Standard Normal DGP (Units: percent)

Model	2	3	4	5	6	7
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
$LR_{uc}(25)$	90.9	32.3	25.8	67.9	3.5	6.3
$LR_{cc}(1)$	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
$LR_{cc}(25)$	84.1	23.8	18.3	55.3	3.9	5.4
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
QPSe1(25)	98.7	78.7	71.7	85.2	75.5	70.9
QPSe2	94.0	78.0	64.1	72.7	67.5	68.6
QPSe3	44.3	37.3	51.9	59.1	46.8	46.5
adjusted[c]	57.6	48.5	66.4	73.6	60.1	60.2

[a] The size of the tests is set at 5%.

[b] Each row represents the percentage of simulations for which the alternative model had a higher QPS score than the "true" model; i.e., the percentage of the simulations for which the alternative model was correctly classified as inaccurate.

[c] The adjusted row for QPSe3 removes the simulations for which the QPS value of the "true" model for the third event is zero; i.e., 23.1% of the simulations.

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} | \Omega_t \sim N(0,1)$. Models 2-5 are normal distributions with variances of 0.5, 0.75, 1.25 and 1.5, respectively. Models 6 and 7 are normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.97$ and $\lambda = 0.99$, respectively.

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

Model	2	3	4	5	6	7
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
$LR_{uc}(25)$	35.3	8.4	30.6	18.9	30.6	18.9
$LR_{cc}(1)$	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
$LR_{cc}(25)$	23.2	8.3	19.9	22.4	43.1	48.0
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
QPSe1(25)	77.0	62.6	81.2	74.9	87.0	81.7
QPSe2	71.7	76.2	79.7	80.4	84.0	84.1
QPSe3	46.1	42.8	48.6	49.3	68.2	67.0
adjusted[c]	52.3	48.5	55.1	55.9	74.3	73.1

[a] The size of the tests is set at 5%.

[b] Each row represents the percentage of simulations for which the alternative model had a higher QPS score than the "true" model; i.e., the percentage of the simulations for which the alternative model was correctly classified as inaccurate.

[c] The adjusted row for QPSe3 removes the simulations for which the QPS value of the "true" model for the third event is equal to zero; i.e., 11.8% of the simulations.

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} | \Omega_t \sim t(6)$. Models 2 and 3 are the homoskedastic models with normal distributions of variance of 1.5 and 1, respectively. Models 4 and 5 are the calibrated heteroskedastic models with the normal distribution, and models 6 and 7 are the calibrated heteroskedastic models with the t(6) distribution.

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

	Model 2	3	4	5	6	7
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
$LR_{uc}(25)$	18.3	38.0	30.4	3.3	3.6	10.9
$LR_{cc}(1)$	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
$LR_{cc}(25)$	15.6	33.0	24.1	5.4	7.6	7.3
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
QPSe1(25)	82.2	85.7	81.2	63.1	64.9	65.9
QPSe2	82.7	85.2	85.1	60.4	63.7	64.1
QPSe3	47.8	41.4	51.9	44.4	44.4	65.8
adjusted[c]	57.3	49.7	60.1	53.1	52.8	73.1

[a] The size of the tests is set at 5%.

[b] Each row represents the percent of simulations for which the alternative model had a higher QPS than the "true" model; i.e., the percent of simulations for which the alternative model was correctly classified as inaccurate.

[c] The adjusted row for QPSe3 removes the simulations for which the QPS value of the "true" model is equal to zero; i.e., 19% of the simulations.

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} | \Omega_t \sim N(0, h_{t+1})$. Models 2, 3 and 4 are the homoskedastic models N(0, 1.5), N(0,1) 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.

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

	Model 2	3	4	5	6	7
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
$LR_{uc}(25)$	87.4	98.9	86.5	8.3	9.0	9.4
$LR_{cc}(1)$	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
$LR_{cc}(25)$	88.8	97.9	82.5	46.2	74.5	7.3
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
QPSe1(25)	99.2	99.8	99.6	90.6	91.8	65.9
QPSe2	93.2	96.2	95.6	82.8	83.0	69.9
QPSe3	59.3	62.6	60.1	52.4	52.3	43.0
adjusted[c]	63.0	66.5	64.1	55.6	55.5	45.7

[a] The size of the tests is set at 5%.

[b] Each row represents the percent of simulations for which the alternative model had a higher QPS score than the "true" model; i.e., the percent of simulations for which the alternative model was correctly classified as inaccurate.

[c] The adjusted row for QPSe3 removes the simulations for which the QPS value of the "true" model is equal to zero; i.e., 6% of the simulations.

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} | \Omega_t \sim t\left(h_{t+1}, 6\right)$. Models 2, 3 and 4 are the homoskedastic models N(0,1.5), N(0,1) 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)-normal model with the same parameter values as Model 1.