REGULATORY EVALUATION OF VALUE-AT-RISK MODELS

Jose A. Lopez

Federal Reserve Bank of New York
Research Paper No. 9710

March 1997

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@frbny.sprint.com

Draft date: March 18, 1997

ABSTRACT: Beginning in 1998, commercial banks may determine their regulatory capital requirements for market risk exposure using value-at-risk (VaR) models; i.e., time-series models of the distributions of portfolio returns. Currently, regulators have available three statistical methods for evaluating the accuracy of VaR models: the binomial method, the interval forecast method and the distribution forecast method. These methods test whether the VaR forecasts in question exhibit properties characteristic of accurate VaR forecasts. However, the statistical tests can have low power against alternative models. A new evaluation method, based on proper scoring rules for probability forecasts, is proposed. Simulation results indicate that this method is clearly capable of differentiating among accurate and 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 those of the Federal Reserve Bank of New York or the Federal Reserve System. I thank Beverly Hirtle, Peter Christoffersen, Frank Diebold, Darryl Hendricks, Jim O'Brien and Philip Strahan as well as participants at the 1996 Federal Reserve System Conference on Financial Structure and Regulation and the Wharton Financial Institutions Center Conference on Risk Management in Banking 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 econometric modeling of financial time series is of obvious interest to financial institutions, whose profits are directly or indirectly tied to their behavior. This exposure is commonly referred to as "market risk". Over the past decade, financial institutions have significantly increased their use of such time series models in response to 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 "market risk" amendment to the 1988 Basle Capital Accord, which proposes U.S. commercial banks with significant trading activities be assessed a capital charge for their "market risk" exposure.[1] Under this amendment to U.S. banking regulations, such regulatory capital charges can now be based on the value-at-risk (VaR) estimates generated by banks' internal VaR models. 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 distribution of portfolio returns.

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

---

[1] For a thorough discussion of the 1988 Basle Capital Accord, see Wagster (1996). For a discussion of the regulatory capital requirements for securities firms, see Dimsom and Marsh (1995).

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 they exhibit correct conditional coverage; and the distribution forecast method proposed by Crnkovic and Drachman (1996) examines whether observed empirical quantiles derived from the 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 desired 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 statistical test, a key issue is their power; i.e., their ability to reject the null hypothesis when it is incorrect. If a statistical test exhibits poor power properties, then the probability of misclassifying an inaccurate model as 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 also 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 statistical testing framework, but instead attempts to gauge the accuracy of VaR models using standard forecast evaluation techniques for probability forecasts. That is, the accuracy of a particular VaR model is gauged by how well forecasts from this model minimize a loss function that represents the regulator's interests. The VaR forecasts used in this evaluation method are probability forecasts of a specified regulatory event, and the loss function used is the quadratic probability score, a proper scoring rule. Although statistical power is not relevant within this framework, the issues of misclassification and comparative accuracy of VaR models under the specified loss function are examined within the context of a Monte Carlo simulation exercise.

The simulation results presented indicate that the three statistical methods can have relatively low power against several alternative hypotheses based on inaccurate VaR models, thus implying that the chances of misclassifying inaccurate models as "acceptably accurate" can be quite high. With respect to the fourth method, the simulation results indicate that the chosen forecast evaluation techniques are capable of distinguishing between accurate 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 estimates as well as the four evaluation methods examined. Sections III and IV outline the simulation experiment and present the results, respectively. Section V summarizes and discusses directions for future research.

II. Evaluating VaR Models

Currently, the most commonly used type of VaR forecasts is VaR estimates. As mentioned above, VaR estimates correspond to a specified quantile of a portfolio's potential loss distribution over a given holding period. To fix notation, let y_t represent portfolio value, which is modeled as the sum of a deterministic component d_t and an innovation ε_t that has a distribution f_t; that is, $y_t = d_t + \varepsilon_t$, where $\varepsilon_t \sim f_t$. The VaR estimate for time t derived from model m conditional on the information available at time t-k, denoted $VaR_{mt}(k,\alpha)$, is the forecasted critical value of f_{mt}, model m's assumed or estimated innovation distribution, that corresponds to its lower α percent tail; that is, $VaR_{mt}(k,\alpha)$ is the solution to

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

Given their roles as internal risk management tools and now as regulatory capital

measures, the evaluation of VaR estimates and the models generating them is of particular interest to both banks and their regulators. Note, however, that the regulatory evaluation of such models differs from institutional evaluations in three important ways. First, the regulatory evaluation has in mind the goal of assuring adequate capital to prevent significant 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 the basic components of the model and their implementation as well as the originating institution can. Third, regulators have the responsibility of constructing evaluations applicable across many 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, commonly known as the "internal models" approach, as well as three statistical evaluation methods are discussed.[2] These methods are based on testing the null hypothesis that the VaR forecasts in question exhibit specified properties characteristic of accurate VaR forecasts. In addition, an alternative method based on comparing probability forecasts of regulatory events of interest with the occurrence of these events is proposed. This method gauges the accuracy of VaR models using a proper scoring rule chosen to match as closely as possible the interests of the banking regulators.

A. Current Regulatory Framework

The current U.S. regulatory framework for the "market risk" exposure of commercial banks' trading accounts is based on an amendment to the 1988 Basle Capital Accord.[3] Beginning

---

[2] Another evaluation method, known as "historical simulation", has been proposed and is based on comparing VaR estimates to a histogram of observed ε_t's. However, as noted by Kupiec (1995), this procedure is highly dependent on the assumption of stationary processes and is subject to the large sampling error associated with quantile estimation, especially in the lower tail of the distribution. This method is not considered here.

[3] This capital requirement covers all positions 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. The final rule applies to

in 1998, regulatory capital charges for "market risk" exposure can be calculated in one of two ways.[4] The first approach, known as the "standardized" approach, consists of regulatory rules that assign capital charges to specific assets and roughly account for selected portfolio effects on banks' risk exposures. However, as reviewed by Kupiec and O'Brien (1995a), this approach has a number of shortcomings with respect to standard risk management procedures.

Under the alternative "internal models" approach, capital requirements are based on the VaR estimates generated by banks' internal risk measurement models using the standardizing regulatory parameters of a ten-day holding period (k = 10) and 99% coverage (α = 1). Thus, a bank's market risk capital is set according to 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 requirement at time t, MRC_{mt}, is based on the larger of $VaR_{mt}(10,1)$ or a multiple of the average of $\left\{ VaR_{mt}(10,1) \right\}_{t=t-60}^{t-1}$; that is,

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

where S_{mt} and SR_{mt} are a regulatory multiplication factor and an additional capital charge for the portfolio's idiosyncratic risk, respectively.

The S_{mt} multiplier links the accuracy of the VaR model to the capital charge by varying over time as a function of the accuracy of the VaR estimates. In the current evaluation framework, S_{mt} is set according to the accuracy of the VaR estimates for a one-day holding period (k = 1) and 99% coverage level (α = 1); thus, an institution must compare its one-day VaR

---

any bank or bank holding company where trading activity equals greater than 10 percent of its total assets or whose trading activity equals greater than $1 billion.

[4] An alternative approach for determining capital charges for banks' "market risk" exposure is the "precommitment" approach that has been proposed by the Federal Reserve Board of Governors; see Kupiec and O'Brien (1995b) for a detailed description.

estimate with the following day's trading outcome.[5] The value of S_{mt} depends on the number of times that daily trading losses exceed the corresponding VaR estimates over the last 250 trading days. Recognizing that even accurate models may perform poorly on occasion and to address the low power of the underlying binomial statistical 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 3, 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 measurement and management system.

Clearly, banking regulators have shifted the emphasis of the "market risk" capital rules toward VaR models. This change in focus necessitates a change in regulatory procedure; specifically, regulators must evaluate the accuracy of the VaR models used to set these new capital requirements.

B. Alternative Evaluation Methods

In this section, four evaluation methods for gauging VaR model accuracy are discussed. For the purposes of this paper and in accordance with the current regulatory framework, the holding period k is set to one. Thus, given a set of one-step-ahead VaR forecasts generated by model m, regulators must determine whether the underlying model is "acceptably accurate", as discussed above. Three statistical evaluation methods using different types of VaR forecasts are available; specifically, evaluation based on the binomial distribution, interval forecast evaluation as proposed by Christoffersen (1997) and distribution forecast evaluation as proposed by

[5] An important question that requires further attention is whether trading outcomes should be defined as the changes in portfolio value that would occur if end-of-day positions remained unchanged (with no intraday trading or fee income) or as actual trading profits. In this paper, the former definition is used.

Crnkovic and Drachman (1996). The underlying premise of these methods is to determine whether the VaR forecasts in question exhibit a specified property of accurate VaR forecasts using a hypothesis testing framework.

However, as noted by Diebold and Lopez (1996), most forecast evaluations are conducted on forecasts that are generally known to be less than optimal, in which case a hypothesis testing framework may not provide much useful information. In this paper, an alternative evaluation method for VaR models, based on the probability forecasting framework presented by Lopez (1997), is proposed. Within this method, the accuracy of VaR models is evaluated using standard forecast evaluation techniques; i.e., how well the forecasts generated from these models minimize a loss function that reflects the interests of regulators.

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

Under the "internal models" approach, banks will report their specified VaR estimates to the regulators, who observe whether the trading losses are less than or greater than these estimates. Under the assumption that the VaR estimates are independent across time, such observations can be modeled as draws from an independent binomial random variable with a probability of exceeding the corresponding VaR estimates equal to the specified α percent.

As discussed by Kupiec (1995), a variety of tests are available to test the null hypothesis that the observed probability of occurrence over a reporting period equals α.[6] The method that regulators have settled on is based on the percentage of exceptions (i.e., occasions where ε_t exceeds $VaR_{mt}(1,\alpha) \equiv VaR_{mt}(\alpha)$) in a sample. The probability of observing x such exceptions in a sample of size T is

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

[6] Kupiec (1995) describes other hypothesis tests that are available and that depend on the bank monitoring scheme chosen by the regulator.

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

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

Note that the LR_{uc} 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. With respect to size, the finite sample distribution for a specific (α,T) pair may be sufficiently different from a $\chi^2_{(1)}$ distribution that the asymptotic critical values might 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 size of the test. As for power, Kupiec (1995) describes how this test has a limited ability to distinguish among alternative hypotheses, even in moderately large samples. Specifically, for sample sizes of regulatory interest (i.e., approximately 250 or 500 trading days) and small values of α, the power of this test in cases where the true α is 110% of the tested α generally does not exceed 10%.

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, the innovation distribution, at a specified probability level α. Given this interpretation, the interval forecast evaluation techniques proposed by Christoffersen (1997) can be applied.[7] The interval forecasts can be evaluated conditionally or unconditionally; that is, forecast performance can be examined over the entire sample period with or without reference to information available at each point in time. The LR_{uc} test is an unconditional test of interval

[7] Interval forecast evaluation techniques are also proposed by Granger, White and Kamstra (1989).

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 for this is that interval forecasts that ignore such variance dynamics might have correct unconditional coverage (i.e., $\alpha^* = \alpha$), but in any given period, may have incorrect conditional coverage; see Figure 1 for an illustration. Thus, the LR_{uc} test does not have power against the alternative hypothesis that the exceptions are clustered in a time-dependent fashion. The LR_{cc} test proposed by Christoffersen (1997) addresses this shortcoming.

For a given coverage level α, one-step-ahead interval forecasts $\left\{\left(-\infty, VaR_{mt}(\alpha)\right)\right\}_{t=1}^{T}$ are generated using model m. From these forecasts and the observed innovations, the indicator variable I_{mt} is constructed as

$$I_{mt} = \begin{cases} 1 \text{ if } \varepsilon_t \in \left(-\infty, VaR_{mt}(\alpha)\right] \\ 0 \text{ if } \varepsilon_t \notin \left(-\infty, VaR_{mt}(\alpha)\right] \end{cases}.$$

Accurate VaR interval forecasts should exhibit the property of correct conditional coverage, which implies that the $\{I_{mt}\}_{t=1}^{T}$ series must exhibit correct unconditional coverage and be serially independent. Christoffersen (1997) shows that the test for correct conditional coverage is formed by combining the tests for correct unconditional coverage and independence as the test statistic $LR_{cc} = LR_{uc} + LR_{ind} \underset{\sim}{a} \chi^2(2)$.

The LR_{ind} statistic is a likelihood ratio statistic of the null hypothesis of serial independence against the alternative of first-order Markov dependence.[8] 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,

[8] Note that 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).

$\pi_{01} = \pi_{11} = \pi$, $L_0 = (1-\pi)^{T_{00}+T_{10}}\pi^{T_{01}+T_{11}}$, and $\pi = (T_{01}+T_{11})/T$. Thus, the test statistic is $LR_{ind} = 2[\log L_A - \log L_0] \overset{a}{\sim} \chi^2(1)$.

B.3. Evaluation of VaR distribution forecasts

Crnkovic and Drachman (1996) state that much of market risk measurement is forecasting f_t, the probability distribution function of the innovation to portfolio value. Thus, they propose to evaluate VaR models based on their forecasted f_{mt} distributions; see Diebold *et al.* (1997) for further discussion. Their evaluation method is based on the observed quantiles, which are the quantiles under $\{f_{mt}\}_{t=1}^{T}$ in which the observed innovations actually fall; i.e., given f_{mt} and the observed ε_t, the corresponding observed quantile is $q_{mt}(\varepsilon_t) = \int_{-\infty}^{\varepsilon_t} f_{mt}(x)dx$. The authors propose to evaluate a VaR model by testing whether observed quantiles derived under the 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 null hypothesis of VaR model accuracy can be tested by determining whether $\{q_{mt}\}_{t=1}^{T}$ are independent and uniformly distributed. (Note that this testing framework permits the evaluation of observed quantiles drawn from possibly time-varying forecasts.)

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 innovation, and the authors suggest the use of the BDS statistic (see Brock *et al.* (1991) to test this hypothesis. However, in this paper, the focus is on their proposed test of the second property.[9] The test of the uniform distribution of $\{p_{mt}\}_{t=1}^{T}$ is based on the Kupier

[9] Note that this emphasis on the second property should understate the power of the overall method since misclassification by this second test might be correctly indicated by the BDS test.

statistic, which measures the deviation between two cumulative distribution functions.[10] Let $D_m(x)$ denote the cumulative distribution function of the observed quantiles, and the Kupier statistic for the deviation of $D_m(x)$ from the uniform distribution is

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

The asymptotic distribution of K_m 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 $G(\lambda) = 2\sum_{j=1}^{\infty}\left(4j^2\lambda^2 - 1\right)e^{-2j^2\lambda^2}$ and $v_m = \max_{0 \le x \le 1}\left|D_m(x) - x\right|$. (Note that for the purposes of this paper, the finite sample distribution of K_m is determined by setting $D_m(x)$ to the true data-generating process in the simulation exercise.) 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 in this paper is based on the probability forecasting framework presented in 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 generated probability forecasts of specified regulatory events minimize a loss function relevant to regulators. The loss functions of interest are drawn from the set of proper probability scoring rules, which can be tailored to the interests of the forecast evaluator. Although statistical power is not relevant within this framework, the degree of model misclassification that characterizes this method can be examined within the context of a

[10] 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.

Monte Carlo simulation exercise.

The proposed evaluation method can be tailored to the interests of the forecast evaluator (in this case, regulatory agencies) in two ways.[11] First, the event of interest to the regulator must be specified.[12] Thus, instead of focussing 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 particular regions of the distributions that are of interest.

In this paper, two types of regulatory events are considered. The first type of event is similar to the one examined above; that is, whether ε_t lies in the lower tail of its distribution. For the purposes of this evaluation method, however, this type of event must be defined differently. Using the unconditional distribution of ε_t based on past observations, the desired empirical quantile loss is determined, and probability forecasts of whether subsequent innovations will be less than it are generated. In mathematical notation, the generated probability forecasts are

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

where $CV(\alpha,F)$ is the lower α% critical value of F, the empirical cumulative distribution function. As currently defined in the "market risk" capital rules, regulators are interested in the lower 1% tail of f_t, but of course, other quantiles might be of interest. The second type of event, instead of focussing on a fixed quantile region of f_{mt}, focusses on a fixed magnitude of portfolio loss. That is, regulators may be interested in determining how well a VaR model can forecast a portfolio loss of p% of y_t over a one-day period. The corresponding probability forecast generated from model m is

---

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

[12] The relevance of such probability forecasts to financial 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."

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

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

The second way of tailoring the 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 generated probability forecasts. The quadratic probability score (QPS), developed by Brier (1950), specifically measures the accuracy of probability forecasts over time and will be used in this simulation exercise. The QPS is the analog of mean squared error for probability forecasts and thus implies a quadratic loss function.[13] The QPS for model m over a sample of size T is

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

where R_t 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.

A key property of the QPS is that it is a strictly proper scoring rule, which means that 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 probability forecast

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

generated by a bank's VaR model, and let $S(r_t, j)$ denote a scoring rule that assigns a numerical score to a probability forecast r_t based on whether the event occurs (j=1) or not (j=0). The reporting bank's expected score is

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

The scoring rule S is strictly proper if $E\left[S\left(P_{mt},j\right) \mid m\right] < E\left[S\left(r_t,j\right) \mid m\right] \forall\ r_t \neq P_{mt}$. Thus, truthful reporting is explicitly encouraged since the bank receives no benefit from modifying their actual forecasts. This property is obviously important in the case of a regulator monitoring and evaluating VaR models that it does not directly observe.[14]

In addition to being an intuitively simple and powerful monitoring tool, QPS highlights the three main attributes of probability forecasts: accuracy, calibration and resolution. As shown by Murphy (1973), the QPS can be decomposed as $QPS = QPS_{\bar{R}} + LSB - RES$, where $QPS_{\bar{R}}$ is QPS evaluated with all the forecasts set equal to the observed frequency of occurrence. Accuracy refers to the closeness, on average, of the predicted probabilities to the observed realizations and is directly measured by QPS. Calibration, which is measured by LSB, refers to the degree of equivalence between the forecasted and observed frequencies of occurrence. Resolution, which is measured by RES, is the degree of correspondence between the average of subsets of the probability forecasts with the average of all the forecasts. Although not used in the following simulation exercise, this decomposition further illustrates the usefulness of the probability forecast framework for tailoring the evaluation of VaR models to the regulator's interests.

The QPS measure is specifically used here because it reflects the regulators' loss function with respect to VaR model evaluation. As outlined in the market-risk regulatory supplement, the goal of reporting VaR estimates is to evaluate the quality and accuracy of a bank's risk

---

[14] The scoring rule S is proper if $E\left[S\left(P_{mt},j\right) \mid m\right] \leq E\left[S\left(r_t,j\right) \mid m\right] \forall\ r_t \neq P_{mt}$. Such scoring rules do not encourage the misreporting of bank's probability forecasts, but they do not guard against it completely.

management system. Since model accuracy is an input into the deterministic capital requirement MRC_t, the regulator should specify a loss function, such as QPS, that explicitly measures accuracy.

III. Simulation Experiment

The simulation experiment conducted in this paper has as its goal an analysis of 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 statistical methods, this amounts to analyzing the power of the statistical tests; i.e., determining the probability with which the tests reject the specified null hypothesis when in fact it is incorrect. With respect to the probability forecasting method, its ability to correctly classify VaR models (i.e., accurate versus inaccurate) 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, equities, interest-sensitive instruments and financial derivatives. However, for the purposes of this exercise, the portfolio in question is simplified to be an integrated process of order one; that is, $y_t = y_{t-1} + \varepsilon_t$, where ε_t has distribution f_t. This specification of y_t, although greatly simplified, can be said to be representative of linear, deterministic conditional mean specifications. It is only for portfolios with nonlinear, deterministic conditional means, such as portfolios with derivative instruments, that this choice presents inference problems.

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 distribution alone. To examine how well the various evaluation methods perform under different distributional assumptions, the

experiments are conducted by setting f_t 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 in ε_t. Specifically, the third segment uses innovations from a GARCH(1,1)-normal process, and the fourth segment uses innovations from a GARCH(1,1)-t(6) process.

In each segment, the true data generating process is one of the seven VaR models evaluated and is designated as the "true" model or model 1. Traditional power analysis of a statistical test is conducted by varying a particular parameter and determining whether the 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 are reasonable alternatives. For example, a popular type of VaR model specifies the variance h_{mt} as an exponentially weighted moving average of squared innovations; that is,

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

This VaR model, as 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.[15] A description of the alternative models used in each segment of the simulation exercise follows.

For the first segment, the true data generating process (DGP) for f_t is the standard normal distribution. 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 is a t(6) distribution. The six alternative models are two

[15] Note that to implement this model, a finite lag-order must be determined. For this exercise, the infinite sum is truncated at 250 observations, which accounts for over 90% of the sum of the weights. See Hendricks (1996) for further discussion on the choice of λ and the truncation lag.

normal distributions with variances of 1 and 1.5 (the same variance as the true f_t) and the two calibrated models with normal distributions as well as with t(6) distributions. For the latter two segments of the exercise, variance dynamics are introduced by using conditional heteroskedasticity of the GARCH form; i.e., $h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}$. In both segments, the true data generating process is a GARCH(1,1) variance process with parameter values $[\omega,\alpha,\beta] = [0.075, 0.10, 0.85]$, which induce an unconditional variance of 1.5. The only difference between the data generating processes of these two segments is the chosen f_t; i.e., the standard normal or the t(6) distribution. The seven models examined in these two segments are the true model; the homoskedastic models of the standard normal, the normal distribution with variance 1.5 and the t-distribution; and the heteroskedastic models of the two calibrated volatility models with normal innovations and the GARCH model with the other distributional form.

In all of the segments, the simulation runs are structured identically. For each run, the simulated y_t series is generated using the chosen data generating process. The length of the in-sample series (after 1000 start-up observations) is set at 2500 observations, which roughly corresponds to ten years of daily observations. The seven alternative VaR models are then used to generate the one-step-ahead VaR forecasts for the next 500 observations of y_t. In the current regulatory framework, the out-of-sample evaluation period is set at 250 observations or roughly one year of daily data, but 500 observations are used in this exercise since the distribution forecast and probability forecast evaluation methods are data-intensive.

The three types of VaR forecasts from the various models are then evaluated using the appropriate evaluation methods. For the binomial and interval forecast methods, the four coverage probabilities examined are $\alpha = [1, 5, 10, 25]$. For the distribution forecast method, only one null hypothesis can be specified. For the probability forecast method, two types of regulatory events are examined. First, using the empirical distribution of ε_t based on the 2500 in-sample observations, $CV(\alpha,F)$, the desired empirical quantile loss, is determined, and probability

forecasts of whether the observed innovations in the out-of-sample period will be less than it are generated.[16] In mathematical notation, these generated probability forecasts are

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

where $CV(\alpha,F)$ is the lower α% critical value of F, the empirical cumulative distribution function of the 2500 observed innovations. The four empirical quantiles examined are α =[1, 5, 10, 25]. Second, a fixed 1% loss of portfolio value is set as the one-day decline of interest, and probability forecasts of whether the observed innovations exceed that percentage loss are generated. Thus,

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

IV. Simulation Results

The simulation results are organized below with respect to the four segments of the simulation exercise; that is, the results for the four evaluation methods are presented for each data generating process and its alternative VaR models. The results are based on 1000 simulations.

Three general points can be made regarding the simulation results. First, the power of the three statistical methods varies considerably against the varied alternative models. 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%). These results indicate that these evaluation methods are likely to misclassify inaccurate models as accurate.

Second, the probability forecasting method seems well capable of determining the accuracy of VaR models. That is, in pairwise comparisons between the true model and an

---

[16] The determination of this empirical quantile of interest is related to, but distinct from, the "historical simulation" approach to VaR model evaluation; see Butler and Schachter (1996).

alternative model, the QPS for the true model is lower than that of 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 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 examined, all four evaluation methods generally seem to be more sensitive to the chosen misspecifications of the distributional shape of f_t than to the chosen misspecifications of the variance dynamics. That is, the four methods seem to be more capable of differentiating between the true model and an alternative model with the same variance dynamics but different distributional shape. 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 test statistics. 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 not significant. However, 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 statistical evaluation methods for a fixed test size of 5%.

- Even though the power results are generally good for the N(0, 0.5) and N(0,1.5) models, overall the statistical tests have only low to moderate power against the chosen alternative models.

- For the LR_{uc} and LR_{cc} test, a distinct asymmetry arises across the homoskedastic normal 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 the relative concentration of the low variance alternatives about the median undermines their tail estimation.

- Both LR tests have no power against the calibrated heteroskedastic alternatives. This result is probably due to the fact that, even though heteroskedasticity is introduced, these alternative models are not very different from the standard normal in the lower tail.

- The K statistic seems to have good power against the homoskedastic models, but low power against the two heteroskedastic models. This result may be largely due to the fact that even though incorrect, these alternative models and their associated empirical quantiles are quite similar to the true model.

Table 2, Panel B contains the five sets of comparative accuracy results for the probability forecast evaluation 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 than that of the alternative model 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. Thus, this method is clearly 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 three statistical evaluation methods for the specified test size of 5%.

- Overall, the power results are low for the LR tests; that is, in the majority of cases, the chosen alternative models are classified as accurate a large percentage of the time.
- However, the K statistic shows significantly higher power against the chosen alternative models. This result seems mainly due to the important differences in the shapes of the alternative models' assumed distributions with respect to the true model.
- With respect to the homoskedastic models, both LR tests generally exhibit good to moderate results for the N(0,1) model, but poor results for the N(0,1.5) model, which has the same variance as the true DGP. With respect to the heteroskedastic models (models 4 through 7), power against these alternatives is generally low with only small differences between the sets of normal and t(6) alternatives.

Table 3, Panel B contains the five sets of comparative accuracy results for the probability forecast evaluation 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 alternative models is greater than that of the true model.

- With respect to the homoskedastic models, this method classifies the N(0,1) model as inaccurate than 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. The reason for this difference is probably that the incorrect form of the variance dynamics more directly affects f_{mt} for the t(6) alternatives (models 6 and 7) more than for the normal alternatives (models 4 and 5).
- With respect to the empirical quantile events, the general pattern is that the distinction between

the true model and the alternative models increases as α increases, but then decreases at α=25. This outcome arises from the countervailing influences of observing more outcomes, which improves model distinction, and movement toward the median, which obscures model distinction. A similar result should be present in the fixed percentage event as a function of the loss percentage p.

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

Table 4, Panel A presents the power analysis of the statistical evaluation methods for the specified test size of 5%. The power results seem to be closely tied to the differences between the distributional shapes of true model and the alternative models.

- With respect to the three homoskedastic VaR models, these statistical methods were able to differentiate between the N(0,1) and t(6) models given the differences between their f_{mt} forecasts and the actual f_t distributions. However, the tests have little power against the N(0,1.5) model, which matches the true model's unconditional variance.
- With respect to the heteroskedastic models, these methods have low power against the calibrated VaR models based on the normal distribution. The result is mainly due to the fact that these smoothed variances are quite similar to the actual variances from the true data-generating process. However, the results for the GARCH-t alternative model vary 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. This result seems to indicate that these statistical tests have little power against close approximations of the variance dynamics but much better power with respect to the distributional assumption of f_{mt}.

Table 4, Panel B presents the five sets of comparative accuracy results for the probability forecast evaluation method. Overall, the results indicate that this method is capable of

differentiating between the true model and alternative models.

- With respect to the homoskedastic models, the loss function is minimized for the true model a high percentage of the time in all five regulatory events, except for the α=1 case for the normal models. In relative terms, the t(6) model is classified as inaccurate more frequently, followed by the N(0,1) model and then the N(0,1.5) model.
- With respect to the heteroskedastic models, the method most clearly distinguishes the GARCH-t model, even though it has the correct dynamics. The two calibrated normal models are only moderately classified as inaccurate. These results seem to indicate that deviations from the true f_t seem to have a greater impact than misspecification of the variance dynamics, especially in the tail.

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

Table 5, Panel A presents the power analysis of the three statistical methods for the specified test size of 5%. The power results seem to be closely tied to the distributional differences between the true model and the alternative models.

- With respect to the homoskedastic models, all three tests have high power; i.e., misclassification is not likely. Specifically, the N(0,1) model, which misspecifies both the variance dynamics and f_t is clearly seen to be inaccurate, although the t(6) model and the N(0,1.5) model are also easily identified as inaccurate.
- With respect to the heteroskedastic models, the LR tests have high power under the α=1 null hypothesis, but this power drops significantly as α increases. The K statistic also has low power against these alternative models. As in the previous segment, these results seem to indicate that the statistical tests have most power against alternative models with misspecified distributional assumptions and less so with respect to model with in accurate variance dynamics.

Table 5, Panel B presents the comparative accuracy results for the probability forecast evaluation method. Once again, the results indicate that the method is capable of differentiating between the true model and the alternative models.

- The comparative results for the regulatory event that ε_t exceeds the lower 1% value of the empirical F distribution are poor while those for the other α-events is much higher. This result is due more to the high volatility and thick tails exhibited by the data-generating process than to the method's ability to differentiate between models. That is, the empirical critical values CV(1,F) were generally so negative as to cause very few observations of the event; so few as to diminish the methods ability to differentiate between the models. However, as α increases, the ability to differentiate between models also increases and becomes quite high.
- With respect to the homoskedastic alternatives, the method is able to accurately classify the alternative models a very high percentage of the time; thus, indicating that incorrect modeling of the variance dynamics can be well detected using this evaluation method.
- With respect to the heteroskedastic alternatives, the method is able to correctly classify the alternative models a moderate to high percentage of the time. Specifically, the calibrated normal models are found to generate losses higher than the true model a high percentage of the time, certainly higher than the GARCH-normal model that captures the dynamics correctly. These results indicate that although approximating or exactly capturing the variance dynamics can lead to a reduction in misclassification, the differences in f_t are still the dominant factor in differentiating between models.

V. Summary

This paper addresses the question of how regulators should evaluate the accuracy of VaR models. The evaluation methods proposed to date are based on statistical hypothesis testing; 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 null hypothesis of model accuracy at the specified significance level. Although such a testing framework can provide useful insight, it hinges on the tests' statistical power; that is, their ability to reject the null hypothesis of model accuracy when the model is inaccurate. As discussed by Kupiec (1995) and as shown in the results contained in this paper, these tests seem to have low power against many reasonable alternatives and thus could lead to a high degree of model misclassification.

An alternative evaluation method, based on the probability forecast framework discussed by Lopez (1997), is proposed and examined. By avoiding hypothesis testing and instead relying on standard forecast evaluation tools, this method attempts to gauge the accuracy of VaR models by determining how well they minimize the loss function chosen by the regulators. The simulation results indicate that this method can distinguish between VaR models; that is, the probability forecasting method seems to be less prone to model misclassification. In addition, it generally seems to be 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.

Butler, J.S. and Schachter, B., 1996. "Improving Value-at-Risk Estimates by Combining Kernel Estimation with Historical Simulation," Manuscript, Vanderbilt University.

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., 1996. "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.

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

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., 1995. "Evaluation of Value-at-Risk Models Using Historical Data," *Federal Reserve Bank of New York Economic Policy Review*, 2, 39-69.

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., 1995a. "The Use of Bank Measurement Models for Regulatory Capital Purposes," FEDS Working Paper #95-11, Federal Reserve Board of Governors.

Kupiec, P. and O'Brien, J.M., 1995b. "A Pre-Commitment Approach to Capital Requirements for Market Risk," Manuscript, Division of Research and Statistics, 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.

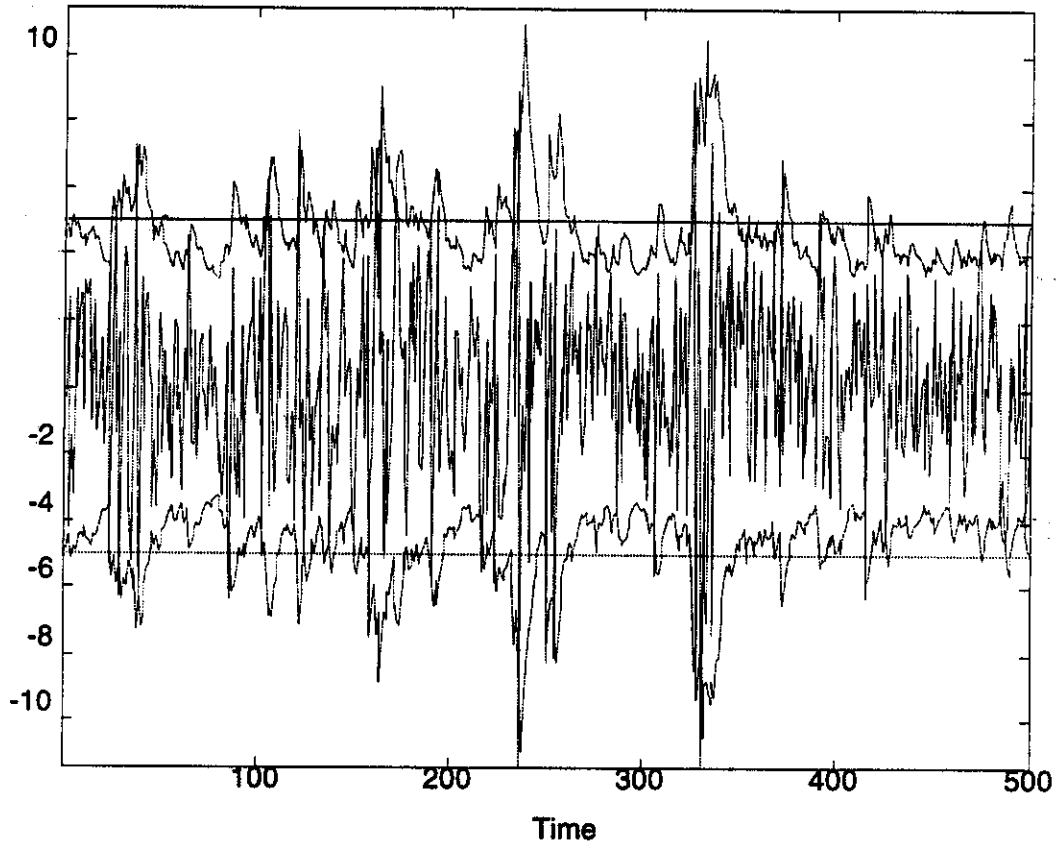
Murphy, A.H., 1973. "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12, 595-600.

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 Uniersity 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) Realization with One-Step-Ahead
90% 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 90% confidence intervals. The straight lines are unconditional confidence intervals, and the jagged lines are conditional confidence intervals based on the true data-generating process. Although both exhibit correct unconditional coverage ($\alpha^*=\alpha=10\%$), only the GARCH confidence intervals exhibit correct conditional coverage.

Table 1. Finite-Sample Critical Values of LR_{uc}, LR_{cc} and K Statistics

	1%	5%	10%
Asymptotic $\chi^2(1)$	6.635	3.842	2.706
$LR_{uc}(99)$	7.111 (1.2%)	4.813 (7.5%)	2.613 (7.5%)
$LR_{uc}(95)$	7.299 (1.2%)	3.888 (6.3%)	3.022 (11.5%)
$LR_{uc}(90)$	7.210 (1.3%)	4.090 (6.2%)	2.887 (11.4%)
$LR_{uc}(75)$	6.914 (1.1%)	3.993 (5.1%)	2.815 (10.2%)
Asymptotic $\chi^2(2)$	9.210	5.992	4.605
$LR_{cc}(99)$	9.702 (1.1%)	4.801 (1.8%)	4.117 (7.0%)
$LR_{cc}(95)$	9.093 (1.0%)	5.773 (4.7%)	4.628 (10.0%)
$LR_{cc}(90)$	9.966 (1.8%)	6.261 (5.6%)	4.768 (11.3%)
$LR_{cc}(75)$	9.541 (1.2%)	6.254 (5.7%)	4.741 (10.7%)
K	0.0800	0.0700	0.0640

The finite-sample critical values are based on a minimum of 1000 simulations. The percentages in parentheses in the panels for the LR tests are the quantiles that correspond to the asymptotic critical values under the finite-sample distributions.

Table 2. Simulation Results for Exercise Segment 1 (Units: percent)

Model	2	3	4	5	6	7
Panel A. Power of the LR_{uc}, LR_{cc} and K Tests Against Alternative VaR Models[a]						
LR_{uc}(99)	99.9	54.6	32.3	70.0	3.3	6.5
LR_{uc}(95)	99.9	68.3	51.5	94.2	2.7	9.2
LR_{uc}(90)	99.9	61.5	47.4	93.1	2.3	7.3
LR_{uc}(75)	90.9	32.3	25.8	67.9	3.5	6.3
LR_{cc}(99)	99.9	56.5	33.1	70.3	4.2	7.9
LR_{cc}(95)	99.9	64.2	40.4	89.2	3.2	9.3
LR_{cc}(90)	99.8	53.0	36.7	86.5	3.2	6.8
LR_{cc}(75)	84.1	23.9	18.3	55.2	3.9	5.5
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(99)	86.4	76.5	83.1	97.2	78.3	66.1
QPSe1(95)	98.9	84.4	82.5	97.9	80.5	74.3
QPSe1(90)	99.6	89.5	82.9	95.3	81.2	76.6
QPSe1(75)	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

[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.

The results are based on 1000 simulations. Model 1 is the true data generating process, 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 λ = 0.97 and λ = 0.99, respectively.

Table 3. Simulation Results for Exercise Segment 2 (Units: percent)

Model	2	3	4	5	6	7
Panel A. Power of the LR_{uc}, LR_{cc} and K Tests Against Alternative VaR Models[a]						
$LR_{uc}(99)$	13.0	86.9	19.6	25.3	21.2	18.1
$LR_{uc}(95)$	11.5	62.1	3.8	3.1	68.1	52.7
$LR_{uc}(90)$	25.7	35.5	13.9	8.0	73.9	60.0
$LR_{uc}(75)$	35.3	8.4	30.6	18.9	30.6	18.9
$LR_{cc}(99)$	15.5	86.1	20.7	28.1	21.3	18.5
$LR_{cc}(95)$	5.9	57.1	2.2	3.9	45.6	32.7
$LR_{cc}(90)$	18.2	29.3	9.2	6.1	61.8	46.6
$LR_{cc}(75)$	24.8	8.4	19.3	12.2	43.0	28.6
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(99)	68.1	84.9	79.1	76.6	96.3	91.0
QPSe1(95)	64.5	88.4	90.5	79.0	98.2	95.2
QPSe1(90)	76.6	79.2	90.0	80.9	97.2	94.2
QPSe1(75)	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

[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.

The results are based on 1000 simulations. Model 1 is the true data generating process, 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 Exercise Segment 3 (Units: percent)

	Model 2	3	4	5	6	7
Panel A. Power of the LR_{uc}, LR_{cc} and K Tests Against Alternative VaR Models[a]						
$LR_{uc}(99)$	22.7	73.9	71.3	4.3	4.8	91.6
$LR_{uc}(95)$	30.7	73.9	72.0	5.4	6.0	81.7
$LR_{uc}(90)$	29.0	65.7	60.3	5.2	5.7	50.0
$LR_{uc}(75)$	18.3	38.0	30.4	3.3	3.6	10.9
$LR_{cc}(99)$	29.3	77.1	73.0	6.4	7.9	91.5
$LR_{cc}(95)$	32.0	72.8	69.3	5.6	6.2	68.6
$LR_{cc}(90)$	30.0	63.1	60.9	5.3	6.2	39.4
$LR_{cc}(75)$	15.3	32.9	24.5	5.2	5.5	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(99)	60.7	66.8	79.2	50.1	51.0	93.0
QPSe1(95)	89.0	92.1	86.4	64.0	66.5	88.8
QPSe1(90)	88.9	93.3	89.9	61.6	66.1	77.1
QPSe1(75)	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

[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.

The results are based on 1000 simulations. Model 1 is the true data generating process, GARCH(1,1)-normal. 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 Exercise Segment 4 (Units: percent)

	Model 2	3	4	5	6	7
Panel A. Power of the LR_{uc}, LR_{cc} and K Tests Against Alternative VaR Models[a]						
$LR_{uc}(99)$	60.8	100.0	96.4	85.8	87.1	86.5
$LR_{uc}(95)$	75.5	100.0	96.9	60.3	63.2	62.1
$LR_{uc}(90)$	80.4	100.0	96.0	36.8	38.5	39.3
$LR_{uc}(75)$	87.4	98.9	86.5	8.3	9.0	9.4
$LR_{cc}(99)$	64.5	100.0	96.7	87.4	89.0	87.7
$LR_{cc}(95)$	82.9	100.0	96.9	56.9	60.9	59.4
$LR_{cc}(90)$	90.1	100.0	96.0	29.4	33.1	29.4
$LR_{cc}(75)$	89.6	98.0	83.1	6.5	6.6	7.8
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(99)	60.7	49.3	49.3	46.3	46.7	41.7
QPSe1(95)	99.6	91.8	90.8	84.2	84.0	69.9
QPSe1(90)	100.0	98.6	98.2	90.4	90.6	76.4
QPSe1(75)	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

[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.

The results are based on 1000 simulations. Model 1 is the true data generating process, GARCH(1,1)-t(6). 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.

FEDERAL RESERVE BANK OF NEW YORK
RESEARCH PAPERS
1997

The following papers were written by economists at the Federal Reserve Bank of New York either alone or in collaboration with outside economists. Single copies of up to six papers are available upon request from the Public Information Department, Federal Reserve Bank of New York, 33 Liberty Street, New York, NY 10045-0001 (212) 720-6134.

9701. Chakravarty, Sugato, and Asani Sarkar. "Traders' Broker Choice, Market Liquidity, and Market Structure." January 1997.

9702. Park, Sangkyun. "Option Value of Credit Lines as an Explanation of High Credit Card Rates." February 1997.

9703. Antzoulatos, Angelos. "On the Determinants and Resilience of Bond Flows to LDCs, 1990 - 1995: Evidence from Argentina, Brazil, and Mexico." February 1997.

9704. Higgins, Matthew, and Carol Osler. "Asset Market Hangovers and Economic Growth." February 1997.

9705. Chakravarty, Sugato, and Asani Sarkar. "Can Competition between Brokers Mitigate Agency Conflicts with Their Customers?" February 1997.

9706. Fleming, Michael, and Eli Remolona. "What Moves the Bond Market?" February 1997.

9707. Laubach, Thomas, and Adam Posen. "Disciplined Discretion: The German and Swiss Monetary Targeting Frameworks in Operation." March 1997.

9708. Bram, Jason, and Sydney Ludvigson. "Does Consumer Confidence Forecast Household Expenditure: A Sentiment Index Horse Race." March 1997.

9709. Demsetz, Rebecca, Marc Saidenberg, and Philip Strahan. "Agency Problems and Risk-Taking at Banks." March 1997.

To obtain more information about the Bank's *Research Papers* series and other publications and papers, visit our site on the World Wide Web (http://www.ny.frb.org). From the research publications page, you can view abstracts for *Research Papers* and *Staff Reports* and order the full-length, hard copy versions of them electronically. Interested readers can also view, download, and print any edition in the *Current Issues in Economics and Finance* series, as well as articles from the *Economic Policy Review.*