METHODS FOR EVALUATING VALUE-AT-RISK ESTIMATES

Jose A. Lopez

Federal Reserve Bank of New York
Research Paper No. 9802

March 1998

Methods for Evaluating Value-at-Risk Estimates

Jose A. Lopez
Research and Market Analysis Group
Federal Reserve Bank of New York
33 Liberty Street
New York, NY 10045
jose.lopez@ny.frb.org

Draft date: March 10, 1998

ABSTRACT:

Beginning in 1998, U.S. commercial banks with significant trading activities must hold capital against their defined market risk exposure. Under the "internal models" approach embodied in the current regulatory guidelines, this capital charge is a function of banks' own value-at-risk (VaR) estimates. Clearly, the accuracy of these VaR estimates is of concern to both banks and their regulators.

To date, two hypothesis-testing methods for evaluating VaR estimates have been proposed; namely, the binomial and the interval forecast methods. For these tests, the null hypothesis is that the VaR estimates in question exhibit a specified property characteristic of accurate VaR estimates. As shown in a simulation exercise, the tests generally have low power and thus are prone to misclassifying inaccurate VaR estimates as "acceptably accurate".

An alternative evaluation method, based on regulatory loss functions, is proposed. Loss functions that capture regulatory concerns are discussed; specifically, the magnitude loss function that assigns a quadratic numerical score when an observed portfolio loss exceeds its VaR estimate. Simulation results indicate that this method is capable of distinguishing between VaR estimates generated by accurate and alternative VaR models. The additional information provided by this method, as well as its flexibility with respect to the specification of the loss function, make a reasonable case for its use in the regulatory evaluation of VaR estimates.

Methods for Evaluating Value-at-Risk Estimates

ABSTRACT:

Beginning in 1998, U.S. commercial banks with significant trading activities must hold capital against their defined market risk exposure. Under the current regulatory guidelines, this capital charge is a function of banks' own value-at-risk (VaR) estimates. Two hypothesis-testing methods for evaluating VaR estimates have been proposed; namely, the binomial and the interval forecast methods. As shown in a simulation exercise, the tests generally have low power and thus are prone to misclassifying inaccurate VaR estimates as "acceptably accurate". An alternative evaluation method, based on loss functions that capture specific regulatory concerns, is proposed. Simulation results indicate that this method is capable of distinguishing between VaR estimates generated by accurate and alternative VaR models. The additional information provided by this method, as well as its flexibility with respect to the specification of the loss function, make a reasonable case for its use in the regulatory evaluation of VaR estimates.

I. Introduction

In August of 1996, the U.S. bank regulatory agencies adopted the market risk amendment (MRA) to the Basle Capital Accord. The MRA, which became effective in 1998, requires that commercial banks with significant trading activities set aside capital to cover the market risk exposure in their trading accounts.[1] The market risk capital requirements are to be based on the "value-at-risk" (VaR) estimates generated by the banks' own risk management models. In general, VaR models attempt to forecast the time-varying distributions of portfolio returns, and VaR estimates are simply specified lower quantiles of these forecasted distributions. In other words, VaR estimates are forecasts of the maximum portfolio loss that could occur over a given holding period with a specified confidence level.

Given the importance of VaR estimates to banks and now their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. As highlighted by Hendricks and Hirtle (1997),

> "The actual benefits from this information depends crucially on the quality and accuracy of the VaR models on which the estimates are based. To the extent that these models are inaccurate and misstate banks' true risk exposures, then the quality of the information derived from any public disclosure will be degraded. More significantly, inaccurate VaR models or models that do not produce consistent estimates over time will undercut the main benefit of a models-based capital requirement: the closer tie between capital requirements and true risk exposures. Thus, validation of the accuracy of these models is a key concern and challenge for supervisors."

To date, two hypothesis-testing methods for evaluating VaR estimates have been proposed: the binomial method, currently the quantitative standard embodied in the MRA, and the interval forecast method proposed by Christoffersen (1998).[2] For these tests, the null hypothesis is that

---

[1] For the details of the market risk amendment, see Federal Register (1996).

[2] Note that other methods for evaluating VaR models have been proposed, but they focus on other aspects of the models' forecasted distributions. For example, Crnkovic and Drachman (1996) focus on the entire forecasted distribution, and Lopez (1997) focusses on probability forecasts generated from the forecasted distributions.

the VaR estimates in question exhibit a specified property characteristic of accurate VaR estimates. If the null hypothesis is rejected, the VaR estimates do not exhibit the specified property, and the underlying VaR model can be said to be "inaccurate". If the null hypothesis is not rejected, then the model can be said to be "acceptably accurate".

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

In addition, this paper proposes an alternative evaluation method that is not based on a hypothesis testing framework, but instead uses standard forecast evaluation techniques. That is, the accuracy of VaR estimates is gauged by how well they minimize a loss function that represents the evaluator's concerns. In this paper, three loss functions that represent specific regulatory concerns are discussed; specifically, the binomial loss function that assigns a numerical score of one when a VaR estimate is exceeded by the corresponding portfolio return; the zone loss function based on the adjustments to the multiplication factor used in the MRA; and the magnitude loss function that assigns a quadratic numerical score when a VaR estimate is exceeded by the portfolio return. Although statistical power is not relevant for this evaluation method, the related issues of comparative accuracy and model misclassification are examined within the context of a simulation exercise.

The simulation results indicate that the hypothesis-testing methods can have relatively low power and thus a relatively high probability of misclassifying inaccurate VaR estimates as "acceptably accurate". For the proposed evaluation method, the simulation results indicate that the degree of model misclassification generally mirrors that of the other methods. However, in certain cases, it provides additional useful information on the accuracy of VaR estimates. Of the

2

three loss functions examined, the magnitude loss function seems to be more capable of distinguishing between accurate and alternative VaR estimates because it incorporates additional information -- the magnitude of the trading losses -- into the evaluation. The ability to use such additional information, as well as the flexibility with respect to the specification of the loss function, make a reasonable case for the use of the loss function method in the regulatory evaluation of VaR estimates.

Section II below describes the current regulatory environment and the three evaluation methods. Section III presents the simulation results that indicate the usefulness of the proposed evaluation method, particularly using the magnitude loss function. Section IV presents a detailed example of how this method can provide additional information useful in the regulatory evaluation of VaR estimates, and Section V concludes.

II. Alternative Evaluation Methods

VaR models are characterized by their forecasted distributions of k-period-ahead portfolio returns. To fix notation, let y_t denote the log of portfolio value at time t. The k-period-ahead portfolio return is $\varepsilon_{t+k} = y_{t+k} - y_t$.[3] 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}.

VaR estimates are the most common type of forecast generated from VaR models. A VaR estimate is simply a specified quantile of the forecasted return distribution over a given holding period. The VaR estimate at time t derived from model m for a k-period-ahead return, denoted $VaR_{mt}(k,\alpha)$, is the critical value that corresponds to the lower α percent tail of f_{mt+k}. Thus, $VaR_{mt}(k,\alpha) = F_{mt+k}^{-1}(\alpha/100)$, where F_{mt+k}^{-1} is the inverse of the cumulative distribution

---

[3]   The usefulness of modeling portfolio returns, as opposed to the underlying market risk factors, in a VaR framework is discussed in Zangari (1997).

3

function corresponding to f_{mt+k}, or equivalently, $VaR_{mt}(k,\alpha)$ is the solution to

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

A. Current Regulatory Framework

The current, U.S. capital rules for the market risk exposure of commercial banks, effective as of 1998, are explicitly based on VaR estimates. The rules cover all assets in a bank's trading account (i.e., assets carried at their current market value) as well as all foreign exchange and commodity positions wherever located in the bank. Any bank or bank holding company whose trading activity accounts for more than ten percent of its total assets or is more than $1 billion must hold regulatory capital against their market risk exposure. The capital charge is to be calculated using the so-called "internal models" approach.[4]

Under this approach, capital charges are based on VaR estimates generated by banks' internal, risk management models using the standardizing parameters of a ten-day holding period (k = 10) and 99 percent coverage (α = 1). In other words, a bank's market risk capital charge is based on its own estimate of the potential loss that would not be exceeded with one percent certainty over the subsequent two week period. The market risk capital that bank m must hold for time t+1, denoted 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,

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

where S_{mt} and SR_{mt} are a multiplication factor and an additional capital charge for the portfolio's

---

[4] Note that an alternative method for monitoring the market risk exposure of commercial banks is the "precommitment" approach proposed by Kupiec and O'Brien (1995).

idiosyncratic credit risk, respectively. Note that, under the current framework, $S_{mt} \geq 3$.

The S_{mt} multiplier is included in the calculation of MRC_{mt+1} for two reasons. First, as described by Hendricks and Hirtle (1997), it adjusts the reported VaR estimates up to what regulators consider to be a minimum capital requirement reflecting their concerns regarding prudent capital standards and model accuracy.[5] Second, S_{mt} explicitly links the accuracy of a bank's VaR model to its capital charge by varying over time. S_{mt} is set according to the accuracy of model m's VaR estimates for a one-day holding period (k = 1) and 99 percent coverage, denoted $VaR_{mt}(1,1)$ or simply VaR_{mt}.

S_{mt} is a step function that depends on the number of exceptions (i.e., occasions when the portfolio return ε_{t+1} is less than VaR_{mt}) observed over the last 250 trading days.[6] The possible number of exceptions is divided into three zones. Within the green zone of four or fewer exceptions, a VaR model is deemed "acceptably accurate", and S_{mt} remains at its minimum value of three. Within the yellow zone of five to nine exceptions, S_{mt} increases incrementally with the number of exceptions. Within the red zone of ten or more exceptions, the VaR model is deemed to be "inaccurate", and S_{mt} increases to its maximum value of four. The institution must also explicitly improve its risk management system.

Clearly, the "internal models" approach embodied in the MRA represents a significant change in how regulatory oversight of bank activities is conducted. Having established that market risk capital will be a function of banks' own VaR estimates, the regulators must now focus on evaluating the accuracy of these VaR estimates. In the following section, three methods for evaluating VaR estimates are discussed. In accordance with the current regulatory framework, one-step-ahead VaR estimates are analyzed.

---

[5] See Stahl (1997) for a mathematical justification of the multiplication factor.

[6] See Finger (1996) for a discussion and example of such exceptions with respect to emerging markets data.

B. Alternative Evaluation Methods

Under the MRA, regulators must determine whether a bank's VaR model is "acceptably accurate" given 250 VaR estimates and the corresponding portfolio returns. To date, two hypothesis-testing methods have been proposed for this type of evaluation: evaluation based on the binomial distribution and interval forecast evaluation, as proposed by Christoffersen (1998). Both methods use hypothesis tests to determine whether the VaR estimates exhibit a specified property characteristic of accurate VaR estimates.

However, as noted by Diebold and Lopez (1996), it is unlikely that forecasts from a model will exhibit all the properties of accurate forecasts. Thus, evaluating VaR estimates solely upon whether a specific property is present may yield only limited information regarding their accuracy. In addition, the power of the tests used in the evaluation must also be considered. In this paper, an evaluation method based on determining how well VaR estimates minimize a regulatory loss function is proposed. Clearly, this evaluation method can provide information that is of direct interest to the regulators.

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

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

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

6

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

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

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

The finite sample size and power characteristics of this test are of interest. With respect to size, the finite sample distribution of LR_{uc} for the specified parameters may be sufficiently different from a $\chi^2(1)$ distribution that the asymptotic critical values may be inappropriate. Table 1, Panel A presents the finite-sample critical values as determined via simulation, and meaningful differences between the two distributions are present and must be accounted for when drawing statistical inference. As for the power of this test, Kupiec (1995) describes how this test has a limited ability to distinguish among alternative hypotheses and thus has low power in samples of size 250.

B.2. Evaluation of VaR using the interval forecast method

VaR estimates are also interval forecasts of the lower one percent tail of f_{t+1}, the one-step-ahead return distribution. Interval forecasts can be evaluated conditionally or unconditionally; that is, with or without reference to the information available at each point in time. The LR_{uc} test is an unconditional test since it simply counts exceptions over the entire period. However, in the presence of time-dependent heteroskedasticity, the conditional accuracy of interval forecasts is an important issue. Interval forecasts that ignore such variance dynamics may have correct unconditional coverage, but at any given time, will have incorrect conditional coverage; see Figure 1 for an illustration. In such cases, the LR_{uc} test is of limited use since it will classify inaccurate VaR estimates as "acceptably accurate".

The LR_{cc} test, adapted from the more general test proposed by Christoffersen (1998), is a test of correct conditional coverage. Given a set of VaR estimates, the indicator variable I_{mt+1} is

7

constructed as

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

Since accurate VaR estimates exhibit the property of correct conditional coverage, the I_{mt+1} series must exhibit both correct unconditional coverage and serial independence. The LR_{cc} test is a joint test of these two properties. The relevant test statistic is $LR_{cc} = LR_{uc} + LR_{ind}$, which is distributed $\chi^2(2)$. The finite sample critical values for the regulatory parameter values are shown in Table 1, Panel B.

The LR_{ind} statistic is the likelihood ratio statistic for the null hypothesis of serial independence against the alternative of first-order Markov dependence.[7] 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, $\pi_{01} = \pi_{11} = \pi$, and the relevant likelihood function is $L_0 = (1-\pi)^{T_{00}+T_{10}}\pi^{T_{01}+T_{11}}$, where $\pi = \left(T_{01}+T_{11}\right)/250$. The test statistic LR_{ind} equals $2\left[\log L_A - \log L_0\right]$ and has an asymptotic $\chi^2(1)$ distribution.

B.3. Evaluation of VaR estimates using regulatory loss functions

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

---

[7]   As discussed in Christoffersen (1998), several other forms of dependence, such as second-order Markov dependence, can be specified. For the purposes of this paper, however, first-order Markov dependence is used.

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

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

where f(x,y) and g(x,y) are functions such that $f(x,y) \geq g(x,y)$. The numerical scores are constructed with a negative orientation; i.e., lower values of C_{mt+1} are preferred since exceptions are given higher scores than non-exceptions. Numerical scores are generated for individual VaR estimates, and the score for the complete, regulatory sample is $C_m = \sum_{i=1}^{250} C_{mt+i}$. Under very general conditions, accurate VaR estimates will generate the lowest possible numerical score.[8] Once a loss function is defined and C_m is calculated, a benchmark can be constructed and used to evaluate the performance of a set of VaR_{mt} estimates. Although many regulatory loss functions can be constructed, below are described the three analyzed in this paper.

a. Loss function implied by the binomial method

The loss function implied by the binomial method is

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

Note that the appropriate benchmark is $E[C_{mt+1}] = 0.01$, which for the full sample is $E[C_m] = 2.5$. As before, only the number of exceptions is of interest, and no additional information over that contained in the binomial method is included in this analysis.

b. Loss function analogous to the adjustment schedule for the S_{mt} multiplier

The numerical score assigned to a set of 250 VaR estimates can be generated by assigning

---

[8]   See Diebold, Gunther and Tay (1997) as well as Granger and Pesaran (1996) for further discussion with respect to distribution and probability forecasts, respectively.

a score to each element of the set or by assigning a score based on the entire set. The adjustment to the S_{mt} multiplier embodied in the MRA is based on the entire set of VaR estimates. Phrased in the notation above, the loss function that generates an analogous numerical score is

$$C_{mt+1} = \begin{cases} 0 \text{ if } \varepsilon_{t+1} \geq VaR_{mt} \\ 0 \quad \text{if } \varepsilon_{t+1} < VaR_{mt} \text{ and } 0 < x \leq 4 \\ 0.4/5 \text{ if } \varepsilon_{t+1} < VaR_{mt} \text{ and } x = 5 \\ 0.5/6 \text{ if } \varepsilon_{t+1} < VaR_{mt} \text{ and } x = 6 \\ 0.65/7 \text{ if } \varepsilon_{t+1} < VaR_{mt} \text{ and } x = 7 \\ 0.75/8 \text{ if } \varepsilon_{t+1} < VaR_{mt} \text{ and } x = 8 \\ 0.85/9 \text{ if } \varepsilon_{t+1} < VaR_{mt} \text{ and } x = 9 \\ 1/x \text{ if } \varepsilon_{t+1} < VaR_{mt} \text{ and } x \geq 10 \end{cases} ,$$

where x is the number of exceptions in the sample and the numerical weights are the actual S_{mt} values divided by x.[9] The benchmark for this numerical score is

$$E\left[C_m \right] = E\left[\sum_{i=1}^{250} C_{mt+i} \right] = \sum_{x=0}^{250} Pr(x) * \left(\sum_{i=1}^{250} C_{mt+i} \mid x \right) = 0.05597.$$

Note that this loss function incorporates the regulatory concerns expressed in the S_{mt} multiplier, but like the binomial loss function, it is based only on the number of exceptions in the sample.

c. Loss function that addresses the magnitude of the exceptions

As noted by the Basle Committee on Banking Supervision (1996), the magnitude, as well as the number, of the exceptions are a matter of concern to regulators. As discussed by

---

[9]   As currently constructed, the S_{mt} adjustment schedule does not address VaR estimates that are possibly too conservative; i.e., VaR estimates that lead to a lower than expected number of exceptions. Given the regulatory interest in providing adequate capital against negative outcomes, the absence of such outcomes is not relevant. However, from the perspective of VaR model evaluation, such outcomes might indicate modeling error. This concern could be addressed by modifying the loss function to include a non-zero score when x < 3.

Hendricks (1996), the magnitude of the observed exceptions can be quite large; in that study, the portfolio losses exceed the corresponding VaR estimate by 30 to 40 percent, on average, and, in the extreme cases, by up to 300 percent.

This concern can readily be incorporated into a loss function by introducing a magnitude term. Although several are possible, a quadratic term is used here, such that

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

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

III. Simulation Exercise

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

11

The first step in this simulation exercise is deciding what type of portfolio to analyze. Although VaR models are commonly applied to complicated portfolios of financial assets, the portfolio value y_{t+1} used here is specified as $y_{t+1} = y_t + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \mid \Omega_t \sim f_{t+1}$. This process is representative of linear, deterministic conditional mean specifications. It is only for portfolios with nonlinear elements, such as portfolios with derivative instruments, that this choice presents inference problems; further research along these lines, as by Pritzker (1996), is needed.

The simulation exercise is conducted in four segments. To examine how the evaluation methods perform under different distributional assumptions, f_{t+1} is set to be the standard normal distribution and a t-distribution with six degrees of freedom, which induces fatter tails than the normal, in the first two segments. The second two segments examine the performance of the evaluation methods in the presence of variance dynamics. The third segment models ε_{t+1} as a GARCH(1,1)-normal process, and the fourth segment does so as a GARCH(1,1)-t(6) process.

In each segment, the true DGP is one of eight 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 termed local alternatives. In this study, non-nested, but common, VaR models are used as reasonable "local" alternatives. For example, a common type of VaR model specifies the variance of ε_{t+1} as an exponentially weighted moving average of squared innovations; that is,

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

This VaR model, a version of which is used in the well-known Riskmetrics calculations (see J.P. Morgan, 1996), is calibrated here by setting λ equal to 0.94 or 0.99, which imply a high-degree

12

of persistence in variance.[10] The alternative VaR models used in each segment of the simulation exercise are described in the subsections below.

The simulation runs are structured identically in each of the segments. For each run, the simulated y_{t+1} series is generated using the chosen DGP. After generating an in-sample period of 3500 observations, the chosen VaR models are used to generate one-step-ahead VaR estimates for the next 250 out-of-sample observations of y_{t+1}. The results are based on 1000 simulations. The simulation results are organized below with respect to the four segments of the exercise.

Two general points can be made. First, with the size of the tests set at 5%, the power of the two hypothesis-testing methods varies considerably against the incorrect null hypotheses implied by the alternative VaR 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%). The results indicate that these two methods are likely to misclassify VaR estimates from inaccurate models as "acceptably accurate".

Second, the degree of model misclassification exhibited by the loss function method roughly matches that of the other two methods; i.e., when the hypothesis-testing methods exhibit low power, the loss function method is generally less capable of distinguishing between accurate and inaccurate VaR estimates. However, overall, the loss function method has a moderate to high ability to gauge the accuracy of VaR estimates. Across the three, regulatory loss functions, the results for the magnitude loss function are relatively better, indicating a greater ability to distinguish between models. This result is not surprising given that the magnitude loss function incorporates additional information -- the magnitude of the exceptions -- into the evaluation.

---

[10] Note that this VaR 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.

A. Simulation results for the homoskedastic standard normal DGP

For the first segment, the true DGP is the standard normal; i.e., $\varepsilon_{t+1} \sim N(0,1)$. The seven alternative models examined are normal distributions with variances of ½, ¾, 1¼ and 1½; the two calibrated VaR models with normal distributions; and the historical simulation model. For this last model, the VaR estimates are formed as the lower 1% quantile of the empirical distribution of the 500 previously observed returns.

Table 2, Panel A presents the power analysis of the hypothesis-testing evaluation methods for a fixed test size of 5%. For the homoskedastic alternatives (models 2 through 5), the power results vary considerably and are related to the differences in the variance; i.e., larger differences induce greater relative power. Similar results are seen for the historical simulation model (model 8). With respect to the calibrated models (models 6 and 7), the tests have no power since, even though unnecessary heteroskedasticity is introduced, these VaR estimates are still quite similar to those of the true DGP.

Table 2, Panel B contains the comparative accuracy results for the loss function method using the specified loss functions. The numerical scores for models 4 and 5, which generate conservative VaR estimates and thus fewer exceptions, are never larger than those for the true DGP. This method cannot distinguish between the true DGP and these models, but this result is acceptable given the regulatory viewpoint implicit in the loss functions; i.e., regulators are concerned if not enough capital is held against possible losses, but not if too much capital is held. However, this method clearly can distinguish between the true DGP and the low variance models (models 2 and 3) that consistently generate smaller VaR estimates than necessary. With respect to the calibrated and historical models (models 6 through 8), the degree of misclassification is generally moderate, although the magnitude loss function exhibits relatively better results.

14

B. Simulation results for the homoskedastic t(6) DGP

For the second segment, the true DGP is a t(6) distribution; i.e., $\varepsilon_{t+1} \sim t(6)$. The seven alternative models are two normal distributions with variances of 1 and 1½ (the same variance as the true DGP); the two calibrated models with normal distributions as well as with t(6) distributions; and the historical simulation model.

Table 3, Panel A show that the overall power of the LR tests against these alternative models is low. With the exception of the N(0,1) model (model 2), the power results are below 50%; thus, the alternative VaR estimates are incorrectly classified as "acceptably accurate" a large percentage of the time. This result is mainly due to the similarity of the alternative VaR models to the true DGP. For example, although models 4 through 7 introduce unnecessary heteroskedsaticity, their VaR estimates are similar to the true, but constant, VaR estimates.

Table 3, Panel B contains the results of the loss function evaluation. For the normality-based models (models 2 through 5), the three loss functions have moderate to high ability to distinguish between alternative VaR estimates, with the zone loss function doing worst and the magnitude loss function doing best. However, with respect to models 6 through 8, this method shows a high degree of model misclassification due to the model's similarity to the true DGP.

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

For the last 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 1½. The only difference between the DGP's in these two segments is the chosen distributional form. For the third segment, $\varepsilon_{t+1} \mid \Omega_t \sim N(0, h_{t+1})$, and for the fourth segment, $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1}, 6)$. The seven alternative VaR models examined are the homoskedastic models of the standard normal, N(0,1½) and the t(6) distribution; the historical simulation model; and the calibrated models with normal innovations and the GARCH model

with the other distributional form.

Table 4, Panel A presents the power analysis of the hypothesis-testing methods. The power results are mainly driven by the differences between the distributional assumptions used by the true DGP and the alternative models. Specifically, the tests have low power against the calibrated normal models (models 5 and 6) since their smoothed variances are quite similar to the true GARCH variances. However, the results for the GARCH-t(6) model (model 7) are much better due to the t(6) assumption. Overall, the hypothesis-testing methods seem to have substantially less power against VaR models characterized by close approximations of the true variance dynamics and have better power against models with incorrect distributional assumptions. With respect to the other models, the LR tests generally have low power since they share the DGP's unconditional variance.

The results for the loss function evaluation method, presented in Table 4, Panel B, are similar; that is, this method has a low to moderate ability to distinguish between the true and alternative VaR models. For the heteroskedastic models, the more conservative GARCH-t(6) model (model 7) obviously minimizes the loss functions due to the smaller number of exceptions. For the calibrated normal models (models 5 and 6) and the historical model (model 8), this method generally has a poor ability to correctly classify them. With respect to the homoskedastic models (models 2 through 4), the degree of misclassification is low for the standard normal (model 2), but much higher for the other two models that have the same unconditional variance as the true DGP. Note, as previously mentioned, that the magnitude loss function is relatively more able to correctly classify VaR estimates than the other loss functions.

D. Simulation results for the GARCH(1,1)-t(6) DGP

Table 5, Panel A presents the power analysis of the hypothesis-testing methods. The power results are clearly tied to the presence of heteroskedasticity in the alternative VaR models.

The homoskedastic models (models 2 through 4) are identified as "inaccurate" with very high power since their VaR estimates cannot match the magnitude of the observed returns from the true DGP. However, for the heteroskedastic models (models 5 through 7) and the historical model (model 8), which are more capable of tracking the underlying variance, the power of the tests declines dramatically.

For the loss function method, the results in Table 5, Panel B indicate that the VaR estimates from the true and alternative models, except the historical model (model 8), can be differentiated. For the homoskedastic alternatives (models 2 through 4), this ability is driven mainly by the fact that constant VaR estimates cannot track the actual returns process well. The heteroskedastic models (models 5 though 7) that can adjust over time do better, but can still be identified as inaccurate due to their misspecificed distributional assumptions. For the historical model, the method's ability to distinguish it from the true DGP is diminished. Note again that, of the three loss functions, the magnitude loss function is again most capable of differentiating between the models.

IV. Implementation of the loss function method

The simulation results presented above indicate that the loss function method is generally capable of distinguishing between VaR estimates from the true DGP and alternative models. Although this ability varies, the method can provide information useful for the regulatory evaluation of VaR estimates, particularly when the magnitude loss function is used. This result is not surprising given that it incorporates the additional information on the magnitude of the exceptions into the evaluation. In this section, this evaluation method is made operational by creating a benchmarking process and by illustrating its use in a detailed example.

A. Creating a benchmark for the observed numerical scores

Under the current regulatory framework, regulators observe $\{\varepsilon_{t+i}, VaR_{mt+i-1}\}_{i=1}^{250}$ for bank m and thus can construct, under the magnitude loss function, $C_m = \sum_{i=1}^{250} C_{mt+i}(\varepsilon_{t+i}, VaR_{mt+i-1})$. However, for a realized value C_m^*, aside from the number of exceptions, not much inference on the performance of the underlying VaR estimates is available; i.e., it is unknown whether C_m^* is a "high" or "low" number. Although comparisons could be made cross-sectionally across banks (m = 1,...,M), a more reasonable method for creating a comparative benchmark is to focus on the distribution of C_m, which is a random variable due to the random observed portfolio returns. Since each observation $\varepsilon_{t+1} \mid \Omega_t \sim f_{t+1}$, additional assumptions on the dependence of the observed returns and their distributions must be imposed in order to analyze $f(C_m)$, the distribution of C_m.

An immediate and commonly used assumption is that the observed returns are independent and identically distributed (iid); i.e., $\varepsilon_{t+1} \sim f$. This is quite a strong assumption, especially given the heteroskedasticity often found in financial time series.[11] However, the small sample size of 250 mandated by the MRA allows few other choices. Having made the assumption that the observed returns are iid, their empirical distribution, denoted $\hat{f}(\varepsilon_{t+1})$, can be estimated using a variety of methods. For example, nonparametric methods, such as smoothed kernel density estimators as per Silverman (1986) or unsmoothed bootstrap methods, could be used. Generally, for issues of tractability, parametric methods are commonly used; i.e., a specific distributional form is assumed, and the necessary parameters are estimated from the available data. For example, if the returns are assumed to be normally distributed with zero mean, the variance can be estimated such that $\hat{f}(\varepsilon_{t+1})$ is $N(0,\hat{\sigma}^2)$.

A reasonable alternative to assuming independence is to impose some form of

---

[11] See Kearns and Pagan (1997) for a discussion of the consequences of ignoring the dependence in financial data when drawing inference about the tails of the data's distribution.

dependence on the data. For example, if the returns are assumed to be driven by a GARCH process, the necessary parameters could be estimated and used to specify the return distributions $\hat{f}(\varepsilon_{t+1} \mid \Omega_t)$, which depends on the information set at time t. However, the small sample size will limit the usefulness of such parameter estimates. The calibrated models previously discussed present a reasonable alternative specification of the dependence in the data. The impact of misspecified temporal dynamics on the construction of $f(C_m)$ relative to that of the iid assumption is not known; further research is necessary. In the example that follows, both types of assumptions are used to examine reported VaR estimates.

Once $\hat{f}(\varepsilon_{t+1})$ or $\hat{f}(\varepsilon_{t+1} \mid \Omega_t)$ has been determined, the empirical distribution of the numerical score C_m under the distributional assumptions, denoted $\hat{f}(C_m)$, can be generated since the distribution of the observed returns and the corresponding VaR estimates are now available. For example, if $\varepsilon_{t+1} \sim N(0, \hat{\sigma}^2)$, then the corresponding VaR estimates are $VaR_{\hat{f}t} = -2.32\hat{\sigma}$. If the assumption is that $\varepsilon_{t+1} \sim N(0, \hat{h}_{t+1})$, then $VaR_{\hat{f}t} = -2.32\sqrt{\hat{h}_{t+1}}$. Using this information, $\hat{f}(C_m)$ can then be constructed via simulation by forming say 1000 values of the numerical score C_m, each based on 250 draws from $\hat{f}(\varepsilon_{t+1})$ and the corresponding VaR estimates.[12]

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

---

[12] Note that although a closed form solution for $\hat{f}(C_m)$ should be available if a parametric assumption is made, simulation methods will be used in this paper.

believe is appropriate, say below 80%, then C_m^* is "typical" under the assumptions made on $\hat{f}(\varepsilon_{t+1})$ and under the regulators' preferences. If $\hat{q}_m$ is above the threshold, then C_m^* can be considered atypical given their preferences, and the regulators should take a closer look at the underlying VaR model.

Note that this method for evaluating VaR estimates does not replace the hypothesis testing methods, but instead, provides complementary information, especially regarding the magnitude of the exceptions. In addition, the flexibility of this method permits many other concerns to be incorporated into the analysis via the choice of the loss function. In the example below, how this method might be employed in an actual case is illustrated, and it can be seen that, in certain cases, the loss function method flags important information not captured in the standard binomial analysis.

B. Detailed example

For this detailed example, the performance of three sets of VaR estimates is examined using the three evaluation methods. As will be shown, inference on the accuracy of the VaR estimates based on the loss function method matches that drawn from the hypothesis-testing methods. However, since it incorporates additional information on the magnitude of the exceptions, the loss function method permits further inference of particular interest to regulators.

The underlying returns process is $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1}, 6)$ with $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$. VaR estimates are generated from three VaR models: the true GARCH-t(6) model; the historical simulation model based on a rolling window of the 500 previous observations; and the calibrated normal model with $\lambda = 0.94$. The models are henceforth denoted as the true, historical and calibrated models, respectively. The 1250 generated observations are analyzed over the five, contiguous but non-overlapping periods of 250 observations. The five periods of simulated data and the corresponding VaR estimates are plotted in figures 2 through 6.

20

Table 6 contains the evaluation results for the two hypothesis-testing methods. Panel A reports the number of exceptions in each of the five periods for the three sets of VaR estimates, and Panels B and C report the LR_{uc} and LR_{cc} statistics, respectively. The occasions for which the null hypothesis of correct unconditional or conditional coverage, respectively, are rejected at the 5% significance level are noted. For the true model, both tests do not reject the null hypothesis that the VaR estimates exhibit the specified properties, and the S_{mt} multiplier would remain at three. For the historical model, the number of exceptions is particularly large in the second and third periods, and the corresponding test statistics reject the null hypotheses. In these periods, S_{mt} increases to its maximum value of four. However, for the other time periods, this VaR model is "acceptably accurate", even though the hypothesis tests indicate a problem in the fifth period when no exceptions occurred. For the calibrated model, the null hypotheses are rejected in only one case, and S_{mt} is above three in all but one period.

Table 7, Panel A contains the numerical scores under the magnitude loss function. As mentioned, these scores alone do not provide much useful inference for evaluating the VaR estimates. However, by making assumptions on the distribution of the observed returns, an approximate distribution of the numerical scores can be created and used to provide a benchmark for evaluation. Once the needed distributional assumptions are in place, $\hat{f}(C_m)$ can be generated via simulation.

Since, in this example, the true DGP is known, the actual $f(C_m)$ can be generated. Table 7, Panel B reports the empirical quantiles q_m under $f(C_m)$. Three results are immediately clear. First, the inference drawn from the loss function method generally matches that of the two hypothesis-testing methods; i.e., the $q_m's$ are generally low (below the threshold 80%), except in a few distinct cases. Second, the $q_m's$ for the historical model in the second and third periods are high (above 80%) due to the large number of exceptions. Third, the $q_m's$ for all of the models are high in the fourth period, even though the number of exceptions is low. Recall that the two

21

hypothesis-testing methods indicated that these three sets of VaR estimates were "acceptably accurate".

The reason for the relatively high scores and $q_m's$ in the fourth period can be seen in Figure 5. Observation number 217 is a particularly large negative number; in terms of relative magnitudes, it exceeds the VaR estimates by about 120% for the true model, 50% for the historical model and 144% for the calibrated model. This result clearly indicates the advantages of the loss function evaluation. By incorporating additional information relevant to the regulator into the evaluation, this method can alert the regulator when an extraordinary event, not detectable by the hypothesis-testing methods, has occurred.

In an actual implementation of the loss function evaluation method, the true DGP is not known. Hence, Panels C and D of Table 7 contain the $q_m's$ under two different $\hat{f}(C_m)$. In Panel C, $\hat{f}(C_m)$ is formed under the assumption that $\varepsilon_{t+1} \sim N(0,\hat{\sigma}^2)$. In Panel D, $\hat{f}(C_m \mid \Omega_t)$ is formed under the assumption that $\varepsilon_{t+1} \sim N(0,\hat{h}_{t+1})$, where $\hat{h}_{t+1}$ follows an exponentially- weighted moving average of squared observed returns with a calibration parameter of 0.94.[13] The empirical quantiles under these two assumed distributions are higher than those under the true DGP, which induces a form of Type I error; that is, under these assumed distributions and for a fixed threshold quantile, the observed $C_m^{*'}s$ will indicate more instances of possibly large exceptions than are called for under the true DGP. The reason for this upward bias is that under these distributional assumptions on ε_{t+1}, the expected value of C_{mt+1} conditional on an exception having occurred will be lower than under the true DGP.[14] Thus, when the $C_m^{*'}s$ are compared to

---

[13]   Note that, in forming the $\hat{h}_{t+1}$ series for each simulation run, an initial value $\hat{h}_1$ must be chosen. The results presented in Table 7, Panel D are based on setting $\hat{h}_1$ equal to the estimated variance of the simulated sample. An alternative specification in which $\hat{h}_1 = \varepsilon_1^2$ generates qualitatively similar results.

[14]   Note that this upward bias in the $\hat{q}_m's$ is brought about by distributional assumptions that generate returns that, conditional on being exceptions, are not as negative as those actually observed. If the distributional assumptions were to generate returns that were generally more negative than actually observed, the bias would go in the opposite direction and cause a form of Type II error; i.e., not indicate concerns when they truly may be present. Although such distributional assumptions could be made, the general concern is that observed returns are being

$\hat{f}(C_m)$ and $\hat{f}(C_m \mid \Omega_t)$, they will generally be in a higher quantile than under $f(C_m)$, as implied by the true DGP.

Although this upward bias is present in the $\hat{q}_m's$, useful inference can still be drawn. If the threshold quantile remains at 80%, the previously noted instances are also found to indicate concern under the two assumed distributions.[15] In addition, four new instances arise: the calibrated model in the first, second and fifth periods as well as the true model in the fifth period.

For the calibrated model in the first period depicted in Figure 2, the observed exceptions range from about 9% to 27% more than their stated VaR estimates, which are relatively low compared to the magnitudes cited by Hendricks (1996). Thus, the "high" $\hat{q}_m's$ for these VaR estimates under the two assumed distributions is based more on the number of exceptions than their magnitude. In this case, inference based on the loss function method provides additional detail, but does not change our overall evaluation of the VaR estimates. In the second period, the number of exceptions is still within the yellow zone, but the loss function method highlights that their magnitude, which range from 5% to 45% beyond the observed return, may be a concern.

For the fifth period, the number of exceptions are again acceptable at two, zero and five for the true, historical and calibrated models, respectively. Although the loss function method cannot provide additional information on the historical model due to the lack of exceptions (an acceptable outcome under this regulatory loss function), the $\hat{q}_m's$ for the other two models are between 80% and 90%. The reason for these high $\hat{q}_m's$ is that the exceptions in both cases are relatively large. The true model's two exceptions are both over 50%, and the calibrated model's five exceptions range from 1% to 50%. Thus, even though both models are "acceptably

---

generated from DGP's with fatter, not thinner, tails than empirically observed.

[15] Note that an alternative way to conduct this type of evaluation is to recognize the upward bias imparted by the assumptions and use a higher threshold quantile, say 90%. This route is complicated by the fact that the proper alternative threshold is not readily apparent. It is simpler to set the threshold quantile quite high at 80% and examine the additional cases with care.

accurate" under the MRA guidelines, the loss function method based on these distributional assumptions provides useful, if biased, information on the performance of the VaR estimates.

V. Conclusion

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

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

The loss function method is based on assigning numerical scores to the performance of the VaR estimates under a loss function that reflects the concerns of the regulators. As shown in the simulation exercise, this loss function method can distinguish between VaR estimates from the actual and alternative VaR models. Furthermore, it allows the evaluation to be tailored to specific interests that regulators may have, such as the magnitude of the observed exceptions. Although when implemented, this evaluation method introduces certain biases due to the necessary distributional assumptions, the analytical results provide useful information on the performance of the VaR estimates. Since these three methods provide complementary information, they could all be useful in the regulatory evaluation of VaR estimates.

24

References

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

Christoffersen, P.F., 1998. "Evaluating Interval Forecasts," *International Economic Review*, forthcoming.

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

Diebold, F.X., Gunther, T.A. and Tay, A.S., 1997. "Evaluating Density Forecasts," NBER Technical Working Paper #215.

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.

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

Finger, C., 1996. "Testing RiskMetrics^{TM} Volatility Forecasts on Emerging Markets Data," *RiskMetrics^{TM} Monitor*, Fourth Quarter, 3-19.

Granger, C.W.J. and Pesaran, M.H., 1996. "A Decision-Theoretic Approach to Forecast Evaluation," Manuscript, Departments of Economics, University of California, San Diego and Cambridge University.

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, 1996. *RiskMetrics Technical Document*, Fourth Edition. New York: JP Morgan.

Kearns, P. and Pagan, A., 1997. "Estimating the Density Tail Index for Financial Time Series," *Review of Economics and Statistics*, *, 171-175.

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. "Regulatory Evaluation of Value-at-Risk Models," Federal Reserve Bank of New York Staff Report #33.

Pritsker, M., 1996. "Evaluating Value at Risk Methodologies: Accuracy versus Computational Time," Manuscript, Federal Reserve Board of Governors.

Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

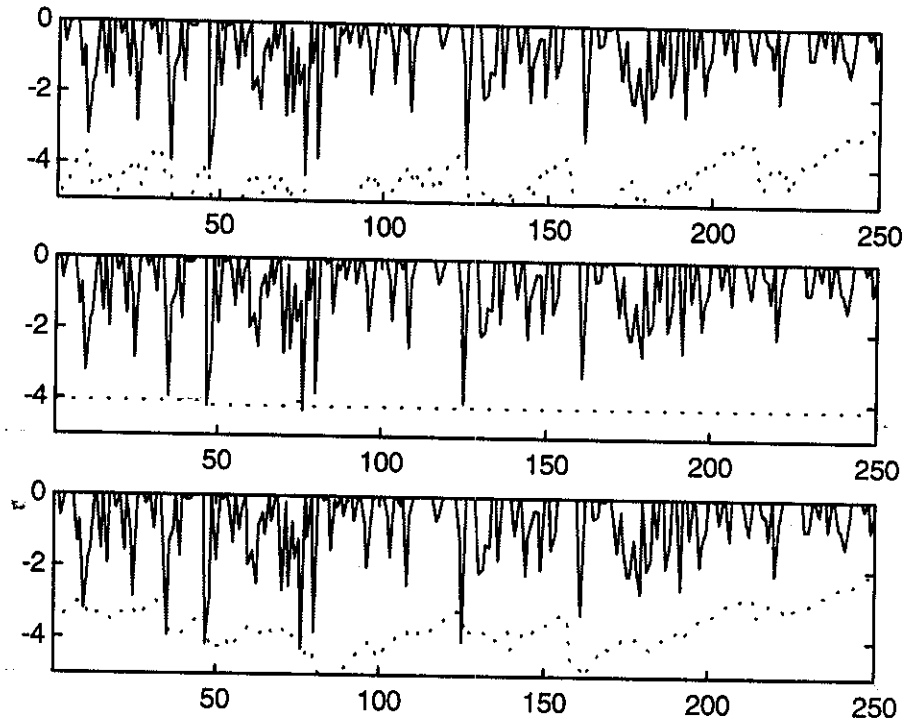Stahl, G., 1997. "Three Cheers," *Risk*, 10, 67-69.

Zangari, P., 1997. "Streamlining the Market Risk Measurement Process," *RiskMetrics*™ *Monitor*, First Quarter, 29-35.

Figure 1.
GARCH(1,1)-Normal Process with One-Step-Ahead,
Lower 5% Conditional and Unconditional Confidence Intervals
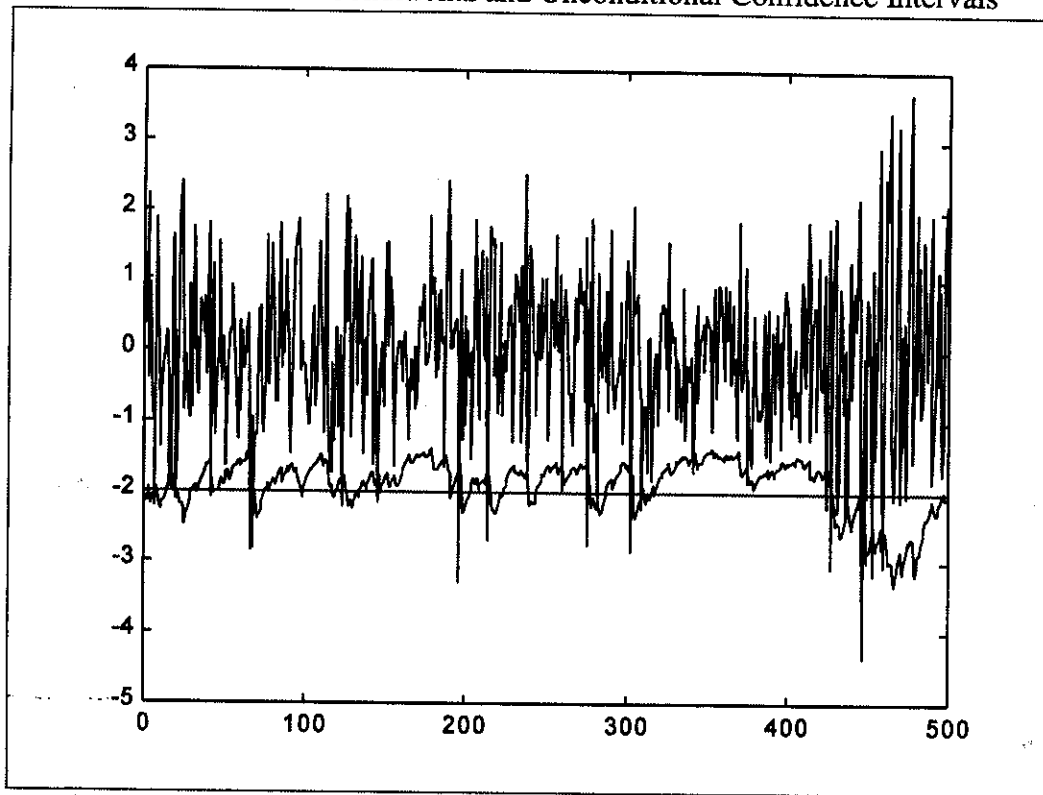
This figure graphs a realization of 500 observations from a GARCH(1,1)-normal process along with two sets of lower 5% confidence intervals. The variance dynamics are characterized as $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$, which implies an unconditional variance of 1.5. The straight line is the unconditional, lower 5% confidence interval based on the unconditional $N(0,1\frac{1}{2})$ distribution, and the jagged line is the conditional, lower 5% confidence intervals based on the true data-generating process. Although both exhibit correct unconditional coverage (i.e., $\alpha^*=0.05$), only the conditional confidence intervals exhibit correct conditional coverage.

The solid line bordering the horizontal zero-axis in each panel represents the observed negative returns. The dotted line in each panel represents the corresponding VaR estimates from each of the three models; i.e., the true model, the historical simulation model and the calibrated model, respectively. The points at which the solid line crosses the dotted line are the exceptions in the sample.

The solid line bordering the horizontal zero-axis in each panel represents the observed negative returns. The dotted line in each panel represents the corresponding VaR estimates from each of the three models; i.e., the true model, the historical simulation model and the calibrated model, respectively. The points at which the solid line crosses the dotted line are the exceptions in the sample.

Figure 4.
Simulated Series for the Illustration:
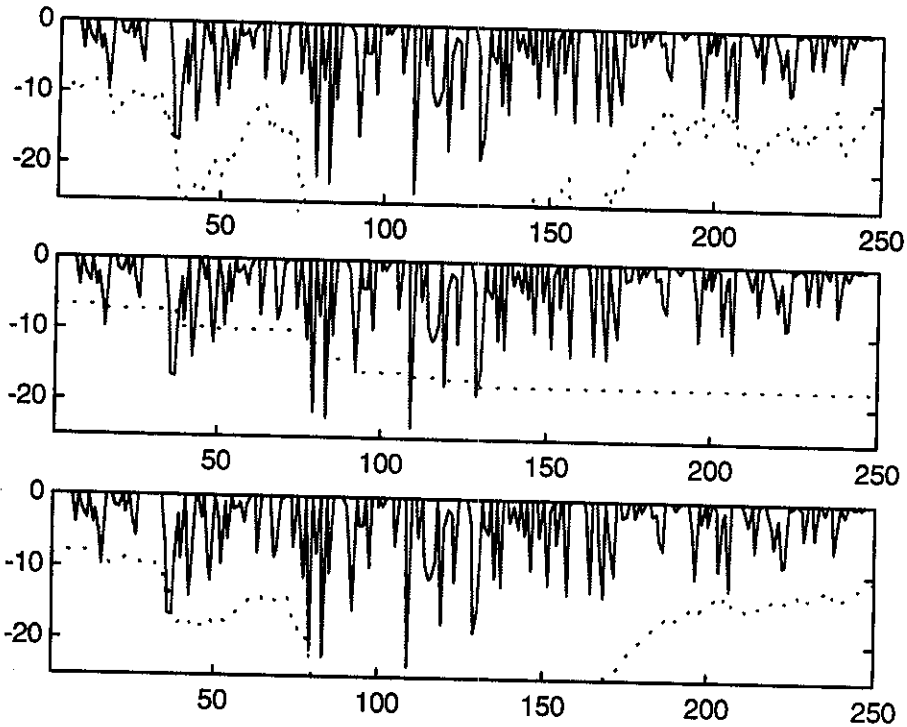**Third Period of 250 Observations**

The solid line bordering the horizontal zero-axis in each panel represents the observed negative returns. The dotted line in each panel represents the corresponding VaR estimates from each of the three models; i.e., the true model, the historical simulation model and the calibrated model, respectively. The points at which the solid line crosses the dotted line are the exceptions in the sample.

Figure 5.
Simulated Series for the Illustration:
Fourth Period of 250 Observations

The solid line bordering the horizontal zero-axis in each panel represents the observed negative returns. The dotted line in each panel represents the corresponding VaR estimates from each of the three models; i.e., the true model, the historical simulation model and the calibrated model, respectively. The points at which the solid line crosses the dotted line are the exceptions in the sample.

Figure 6.
Simulated Series for the Illustration:
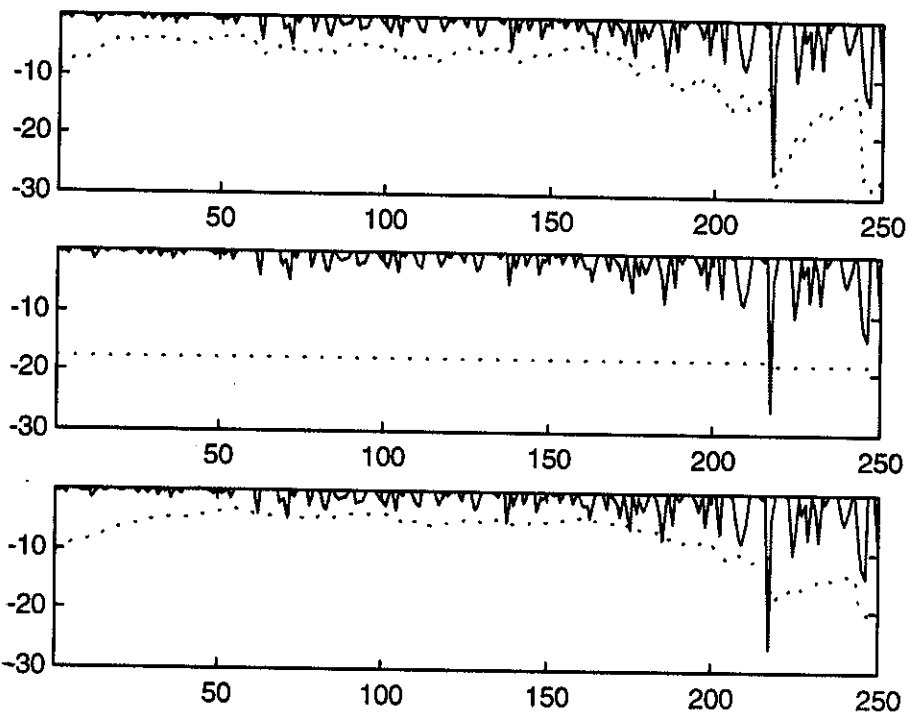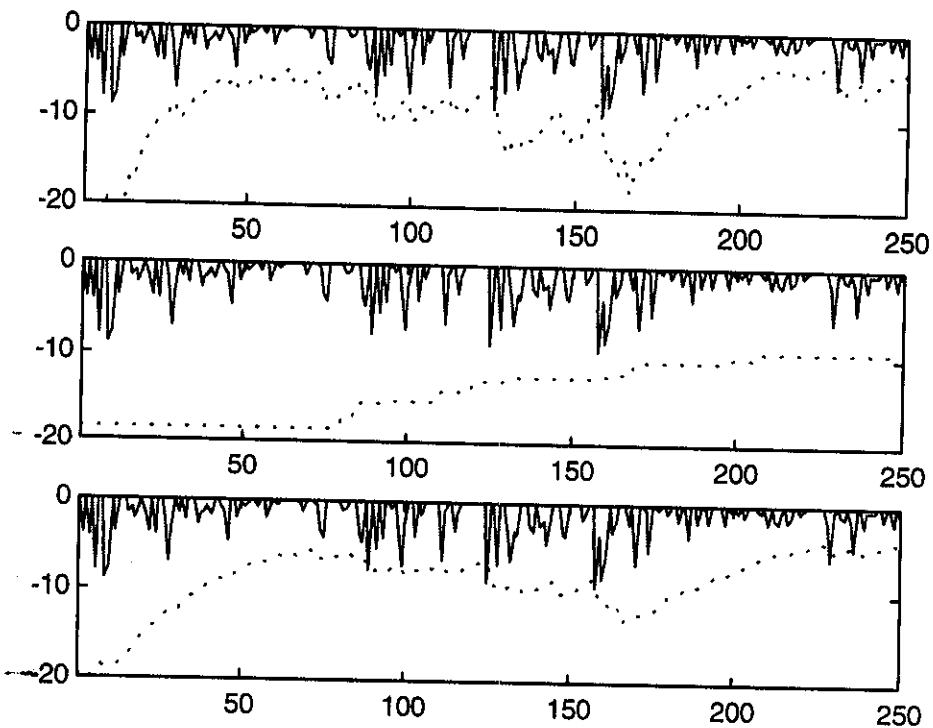**Fifth Period of 250 Observations**

The solid line bordering the horizontal zero-axis in each panel represents the observed negative returns. The dotted line in each panel represents the corresponding VaR estimates from each of the three models; i.e., the true model, the historical simulation model and the calibrated model, respectively. The points at which the solid line crosses the dotted line are the exceptions in the sample.

Table 1. Finite-Sample Critical Values for the LR_{uc} and LR_{cc} Statistics

<u>Significance Level</u>

	<u>1%</u>	<u>5%</u>	<u>10%</u>
Panel A. Critical Values for the LR_{uc} Statistic			
Asymptotic $\chi^2(1)$	6.635	3.842	2.706
Finite-Sample	5.497 (0.5%)	5.025 (9.5%)	3.555 (12.2%)
Panel B. Critical Values for the LR_{cc} Statistic			
Asymptotic $\chi^2(2)$	9.210	5.992	4.605
Finite-Sample	6.007 (0.2%)	5.005 (1.1%)	5.005 (11.8%)

The finite-sample critical values for the LR_{uc} and LR_{cc} test statistics for the lower one percent quantile ($\alpha=1$) are based on 10,000 simulations of sample size T = 250. The percentages in parentheses are the quantiles that correspond to the asymptotic critical values under the finite-sample distribution.

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

Models

	Homoskedastic				Heteroskedastic		Historical
	2	3	4	5	6	7	8

Panel A. Power of the LR_{uc} and LR_{cc} Tests Against Alternative VaR Models [a]

LR_{uc}	97.2	30.4	29.7	54.9	4.3	4.5	40.2
LR_{cc}	97.8	32.9	30.5	60.1	5.4	5.7	43.4

Panel B. Accuracy of VaR Estimates Using Regulatory Loss Functions [b]

Loss function

Binomial	100	94.4	0.0	0.0	55.3	55.4	28.3
Zone	99.6	66.8	0.0	0.0	17.9	18.2	6.7
Magnitude	100	99.7	0.0	0.0	76.1	76.4	53.8

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

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

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \sim N(0,1)$. Models 2 through 5 are homoskedastic normal distributions with variances of ½, ¾, 1¼ and 1½, respectively. Models 6 and 7 are normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.94$ and $\lambda = 0.99$, respectively. Model 8 is the historical simulation model based on the previous 500 observations.

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

<u>Models</u>

	<u>Homoskedastic</u>		<u>Heteroskedastic</u>				<u>Historical</u>
	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>

Panel A. Power of the LR_{uc} and LR_{cc} Against Alternative VaR Models[a]

	2	3	4	5	6	7	8
LR_{uc}	59.1	10.8	15.3	14.6	20.3	19.9	7.9
LR_{cc}	61.5	11.2	17.4	19.9	30.4	30.5	12.4

Panel B. Accuracy of VaR Models Using Regulatory Loss Functions[b]

<u>Loss function</u>	2	3	4	5	6	7	8
Binomial	99.2	69.8	85.5	85.5	5.1	5.0	26.3
Zone	85.0	27.1	47.5	47.3	0.2	0.1	5.4
Magnitude	99.9	97.4	97.3	97.2	10.7	10.3	51.0

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

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

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \sim$ t(6). Models 2 and 3 are the homoskedastic models with normal distributions of variance of 1 and 1.5, 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. Model 8 is the historical simulation model based on the previous 500 observations.

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

<u>Models</u>

	<u>Homoskedastic</u>			<u>Heteroskedastic</u>			<u>Historical</u>
	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>

Panel A. Power of the LR_{uc} and LR_{cc} Against Alternative VaR Models [a]

	2	3	4	5	6	7	8
LR_{uc}	52.3	21.4	30.5	5.1	10.3	81.7	23.2
LR_{cc}	56.3	25.4	38.4	6.7	11.9	91.6	33.1

Panel B. Accuracy of VaR Estimates Using Regulatory Loss Functions [b]

<u>Loss function</u>

	2	3	4	5	6	7	8
Binomial	91.7	41.3	18.1	52.2	48.9	0	38.0
Zone	72.1	21.0	8.1	15.2	18.4	0	17.7
Magnitude	96.5	56.1	29.1	75.3	69.4	0	51.5

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

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

The results are based on 1000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \mid \Omega_t \sim N(0, h_{t+1})$.
Models 2, 3 and 4 are the homoskedastic models N(0, 1), N(0,1.5) and t(6), respectively. Models 5 and 6 are the two calibrated heteroskedastic models with the normal distribution, and model 7 is a GARCH(1,1)-t(6) model with the same parameter values as Model 1. Model 8 is the historical simulation model based on the previous 500 observations.

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

<u>Models</u>

	<u>Homoskedastic</u>			<u>Heteroskedastic</u>			<u>Historical</u>
	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>

Panel A. Power of the LR_{uc} and LR_{cc} Against Alternative VaR Models [a]

LR_{uc}	99.8	97.5	94.4	17.9	34.7	59.1	47.3
LR_{cc}	99.9	97.7	95.6	23.7	35.6	61.5	54.8

Panel B. Accuracy of VaR Estimates Using Regulatory Loss Functions [b]

<u>Loss function</u>

Binomial	99.9	99.9	99.8	82.6	66.9	99.2	42.4
Zone	99.9	99.0	97.1	47.2	42.7	85.0	29.9
Magnitude	99.9	99.9	99.9	94.8	78.0	99.9	53.7

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

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

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), N(0,1.5) and t(6), respectively. Models 5 and 6 are the two calibrated heteroskedastic models with the normal distribution, and model 7 is a GARCH(1,1)-normal model with the same parameter values as Model 1. Model 8 is the historical simulation model based on the previous 500 observations.

Table 6. Hypothesis-Testing Results for the Detailed Example

	Model True	Historical	Calibrated
Panel A. Number of Exceptions			
First period	1	2	5
Second period	3	11	6
Third period	1	14	3
Fourth period	1	1	6
Fifth period	2	0	5
Panel B. LR_{uc} Statistics			
First period	1.1765	0.1084	1.9568
Second period	0.0949	15.8906*	3.5554
Third period	1.1765	25.7803*	0.0949
Fourth period	1.1765	1.1765	3.5554
Fifth period	0.1084	5.0252*	1.9568
Panel C. LR_{cc} Statistics			
First period	1.1846	0.1408	2.1617
Second period	0.1681	16.9078*	3.8517
Third period	1.1846	30.1907*	5.5202*
Fourth period	1.1846	1.1846	3.8517
Fifth period	0.1408	5.0252*	2.1617

The time periods are based on a division of the entire simulation run of 1250 observations into five, contiguous but non-overlapping periods of 250 observations. The true model is $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1}, 6)$ with $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$. The historical simulation model is based on the 500 previous observations. The calibrated model uses the calibrated variance parameter of $\lambda = 0.94$ and the normal distribution.

The asterisk indicates that the null hypothesis is rejected at the 5% significance level using the finite-sample critical values presented in Table 1.

Table 7. Magnitude Loss Function Results for the Detailed Example

	Model True	Historical	Calibrated
Panel A. Numerical Scores			
First period	1.1287	2.0803	7.1048
Second period	3.8180	58.3150	15.8955
Third period	1.4854	507.5814	24.7188
Fourth period	200.1094	71.4351	243.8740
Fifth period	15.6136	0.0	16.9524
Panel B. Empirical Quantiles under the True DGP (Units: Percent)			
First period	13.7	22.0	54.1
Second period	31.0	**89.6**	64.6
Third period	11.3	**86.9**	37.6
Fourth period	**95.9**	**86.1**	**97.0**
Fifth period	53.8	0.0	56.1
Panel C. Empirical Quantiles under the Normal Distribution (Units: Percent)			
First period	17.4	29.0	**88.6**
Second period	44.0	**100.0**	**91.7**
Third period	10.5	**99.8**	46.5
Fourth period	**100.0**	**99.8**	**100.0**
Fifth period	**82.7**	0.0	**84.3**
Panel D. Empirical Quantiles under the Calibrated Normal Distribution (Units: Percent)			
First period	20.5	33.1	**90.1**
Second period	52.9	**99.6**	**93.4**
Third period	13.6	**99.3**	62.2
Fourth period	**99.9**	**98.8**	**99.9**
Fifth period	**86.7**	0.0	**88.5**

The time periods are based on a division of the entire simulation run of 1250 observations into five non-overlapping periods of 250 observations. The true model is $\varepsilon_{t+1} | \Omega_t \sim t(h_{t+1}, 6)$ with $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85 h_t$. The historical simulation model is based on the 500 previous observations. The calibrated model uses the calibrated variance parameter of $\lambda = 0.94$. Empirical quantiles over the selected threshold of 80% are in bold.

FEDERAL RESERVE BANK OF NEW YORK
RESEARCH PAPERS
1998

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.

9801. Higgins, Matthew, and Carol Osler. "Asset Market Hangovers and Economic Growth: U.S. Housing Markets." January 1998.

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/rmaghome). 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*.