Wharton

Financial
Institutions
Center

Regulatory Evaluation of Value-at-Risk Models

by
Jose A. Lopez

96-51

The Wharton School
University of Pennsylvania

THE WHARTON FINANCIAL INSTITUTIONS CENTER

The Wharton Financial Institutions Center provides a multi-disciplinary research approach to the problems and opportunities facing the financial services industry in its search for competitive excellence. The Center's research focuses on the issues related to managing risk at the firm level as well as ways to improve productivity and performance.

The Center fosters the development of a community of faculty, visiting scholars and Ph.D. candidates whose research interests complement and support the mission of the Center. The Center works closely with industry executives and practitioners to ensure that its research is informed by the operating realities and competitive demands facing industry participants as they pursue competitive excellence.

Copies of the working papers summarized here are available from the Center. If you would like to learn more about the Center or become a member of our research community, please let us know of your interest.

Anthony M. Santomero
Director

The Working Paper Series is made possible by a generous grant from the Alfred P. Sloan Foundation

Regulatory Evaluation of Value-at-Risk Models [1]

Draft Date: September 9, 1996

Abstract: Value-at-risk (VaR) models have been accepted by banking regulators as tools for setting capital requirements for market risk exposure. Three statistical methodologies for evaluating the accuracy of such models are examined; specifically, evaluation based on the binomial distribution, interval forecast evaluation as proposed by Christoffersen (1995), and distribution forecast evaluation as proposed by Crnkovic and Drachman (1995). These methodologies test whether the VaR forecasts in question exhibit properties characteristic of accurate VaR forecasts. However, the statistical tests used often have low power against alternative models. A new evaluation methodology, based on the probability forecasting framework discussed by Lopez (1995), is proposed. This methodology gauges the accuracy of VaR models using forecast evaluation techniques. It is argued that this methodology provides users, such as regulatory agencies, with greater flexibility to tailor the evaluations to their particular interests by defining the appropriate loss function. Simulation results indicate that this methodology is clearly capable of differentiating among accurate and alternative VaR models.

Jose Lopez is in the 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

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 as well as Frank Diebold, Darryl Hendricks and Philip Strahan for their comments.

This paper was presented at the Wharton Financial Institutions Center's conference on Risk Management in Banking, October 13-15, 1996.

> 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, 1996)

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. 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 "market risk" supplement to the 1988 Basle Capital Adequacy Accord, which proposes that institutions with significant trading activities be assessed a capital charge for their market risk exposure. Under the proposed "internal models" approach, such regulatory capital requirements would 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.

Given the importance of VaR forecasts to banks and their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. Three statistical evaluation methodologies 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 the binomial distribution, currently the basis of the regulatory supplement and extensively discussed by Kupiec (1995), examines whether VaR estimates exhibit correct unconditional coverage; the interval forecast method proposed by Christoffersen (1995) examines whether they exhibit correct conditional coverage; and the distribution forecast method proposed by Crnkovic and Drachman (1995) examines whether observed empirical percentiles are independent and uniformly distributed. In these tests, if the null hypothesis is rejected, the underlying VaR model is said to be inaccurate, and if not rejected, then the model can be said to be 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 methodology based on the probability forecasting framework presented by Lopez (1995). In contrast to those listed above, this methodology is not based on a statistical testing framework, but instead attempts to gauge the accuracy of VaR models using standard forecast evaluation techniques. That is, a regulatory loss function is specified, and the accuracy of VaR forecasts (and their underlying model) is gauged by how well they minimize this loss function. The VaR forecasts used in this methodology are probability forecasts of a specified regulatory event, and the loss function used is the quadratic probability score (QPS). 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 methodologies can have relatively low power against several alternative hypotheses based on inaccurate VaR models, thus implying that the chances of misclassifying inaccurate models as accurate can be quite high. With respect to the fourth methodology, 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 methodologies examined. Sections III and IV outline the simulation experiment and present the results, respectively. Section V summarizes and discusses directions for future research.

II. The Evaluation of VaR Models

Currently, the most commonly used type of VaR forecasts is VaR estimates. As defined above, VaR estimates correspond to a specified percentile 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 based on the information available at time t-k, $V_{mt}(k,\alpha)$, 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, $V_{mt}(k,\alpha)$ is the solution to

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

Given their roles as internal risk management tools and regulatory capital measures, the evaluation of the models generating VaR estimates is of interest to banks and their regulators. Note, however, 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 as well as the originating institution can. Third, regulators have the responsibility of constructing evaluations applicable across many institutions.

In this section, the current regulatory framework, commonly known as the "internal models" approach, as well as three statistical evaluation methodologies are discussed.[1] These methodologies test the null hypothesis that the VaR forecasts in question exhibit specified properties characteristic of accurate VaR forecasts. In addition, an alternative methodology based on comparing probability forecasts of regulatory events of interest with the occurrence of these events is proposed. This methodology gauges the accuracy of VaR models using a loss function tailored to the interests of the regulatory agencies.

A. Current Regulatory Framework

The current regulatory framework for market risk is based on the general principles set forth in the 1988 Basle Capital Adequacy Accord, which proposed minimum capital requirements for banks' credit risk exposure. In August 1996, American bank regulatory agencies adopted a supplement to the Accord that proposed minimum capital requirements for banks' market risk exposure. The supplement consists of two alternative approaches for setting such capital standards for bank trading accounts, which are bank assets carried at their current market value.[2]

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

---

[1] Another evaluation methodology, 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.

[2] A third approach known as the "precommitment" approach has been proposed by the Federal Reserve Board of Governors; see Kupiec and O'Brien (1995b) for a detailed description.

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 ($\alpha = 1$). 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.

A bank’s market risk capital requirement at time t, MRC_{mt}, is based on a multiple of the larger of $VaR_{mt}(10,1)$ or the average of $\left\{ VaR_{mt}(10,1) \right\}_{t=t-60}^{t-1}$; that is,

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

where S_{mt} and SR_{mt} are a regulatory multiplication factor and an additional capital charge for the portfolio’s specific 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 ($\alpha = 1$); thus, an institution must compare its one-day VaR estimate with the following day’s trading outcome.[3] 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

[3] 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.

deemed to be inaccurate, and S_{mt} increases to four. The institution must also explicitly improve its risk measurement and management system.

B. Alternative Evaluation Methodologies

In this section, four evaluation methodologies 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 "accurate". Three statistical evaluation methodologies using different types of VaR forecasts are available; specifically, evaluation based on the binomial distribution, interval forecast evaluation as proposed by Christoffersen (1995) and distribution forecast evaluation as proposed by Crnkovic and Drachman (1995). The underlying premise of these evaluation methodologies is to determine whether the VaR forecasts 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 methodology for VaR models, based on the probability forecasting framework presented by Lopez (1995), is proposed. Within this methodology, the accuracy of VaR models is evaluated using standard forecast evaluation techniques; i.e., by how well they 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 VaR estimates to the regulators, who observe whether the trading losses are less than or greater than the 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 desired α percent.

As discussed by Kupiec (1995), a variety of tests are available to test the null hypothesis that the observed probability of occurence over a reporting period equals α.[4] The method that regulators have settled on is based on the proportion 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

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

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 must 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 little ability to distinguish among alternative hypotheses, even in moderately large samples.

---

[4] Kupiec (1995) describes several hypothesis tests that are available and depend on how the bank is monitored. However, this paper focusses on daily reporting and evaluation after a fixed number of days.

B. 2. *Evaluation of VaR interval forecasts (Christoffersen, 1995)*

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 (1995) can be applied.[5] 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 forecasts since it ignores this type of information.

However, as argued by Christoffersen (1995), in the presence of the time-dependent, variance dynamics 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 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 (1995) addresses this shortcoming.

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

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

Accurate VaR interval forecasts should exhibit the property of correct conditional coverage, which implies that the $\left\{I_{mt}\right\}_{t=1}^{T}$ series must exhibit correct unconditional coverage and be serially independent. Christoffersen (1995) shows that the test for correct conditional coverage is formed

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

by combining the tests for correct unconditional coverage and independence as the test statistic is $LR_{cc} = LR_{uc} + LR_{ind} \overset{a}{\sim} \chi^2(2)$. The LR_{ind} statistic is a likelihood ratio statistic of the null hypothesis of serial independence against the alternative of first-order Markov dependence.[6] The likelihood function under this alternative hypothesis is $L_A = (1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}}$, where $\pi_{01} = T_{01}/(T_{00}+T_{01})$ and $\pi_{11} = T_{11}/(T_{10}+T_{11})$. Note that the T_{ij} notation denotes the number of observations in state j after having been in state I the period before. Under the null hypothesis of independence, $\pi_{01} = \pi_{11} = \pi$, $L_0 = (1-\pi)^{T_{00}+T_{10}}\pi^{T_{01}+T_{11}}$, and $\pi = (T_{01}+T_{11})/T$. Thus, the proposed LR_{ind} 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, 1995)

Crnkovic and Drachman (1995) 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. Their methodology is based on testing whether observed quantiles derived from $\{f_{mt}\}_{t=1}^{T}$ exhibit the properties of observed percentiles from accurate distribution forecasts. The observed percentiles are the quantiles under $\{f_{mt}\}_{t=1}^{T}$ in which the observed innovations actually fall; i.e., given $f_{mt}(x)$ and the observed ε_t, the corresponding observed percentile is $p_{mt}(\varepsilon_t) = \int_{-\infty}^{\varepsilon_t} f_{mt}(x)dx$. Since the percentiles 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 $\{p_{mt}\}_{t=1}^{T}$ is independent and uniformly distributed. Note that this testing framework allows for the aggregate evaluation of the f_{mt} forecasts, even though they may be time-varying.

Crnkovic and Drachman (1995) suggest that these two properties be examined separately and thus propose two separate hypothesis testing procedures. As in the interval forecast method,

[6] Note that higher-order dependence could be specified. Christoffersen (1995) also presents an alternative test of this null hypothesis based on David (1947).

the independence of the observed percentiles indicates whether the VaR model captures the higher-order dynamics of the innovation, and the authors suggest the use of the BDS statistic to test this hypothesis. However, in this paper, the focus is on their proposed test of the second property.[7] The test of the uniform distribution of $\{p_{mt}\}_{t=1}^{T}$ is based on the Kupier statistic, which measures the deviation between two cumulative distribution functions.[8] Let $D_m(x)$ denote the cumulative distribution function of the observed percentiles, and the Kupier statistic for the deviation of $D_m(x)$ from the uniform distribution is

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

The distribution of K_m is

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

where $G(\lambda) = 2\sum_{j=1}^{\infty}\left(4j^{\lambda^2} - 1\right)e^{-2j^2\lambda^2}$ and $v_m = \max_{0 \le x \le 1}\left|D_m(x) - x\right|$. 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 methodology proposed in this paper is based on the probability forecasting framework presented in Lopez (1995). As opposed to the hypothesis testing methodologies

---

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

[8] Crnkovic and Drachman (1995) 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.

discussed previously, this methodology 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 the relevant loss function. The loss functions of interest are drawn from the set of 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 methodology is examined within the context of a 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.[9] First, the event of interest must be specified Thus, instead of focussing exclusively on a fixed percentile of the forecasted distributions or on the entire distributions themselves, this methodology allows for the evaluation of the VaR models based upon the particular regions of the distributions 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 proposed evaluation methodology, however, this type of event is defined differently. Using the observed unconditional distribution of ε_t, the desired empirical quantile loss is determined, and probability forecasts of whether the 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(α,F) is the lower α% critical value of F, the empirical cumulative distribution function. As currently defined, regulators are interested in the lower 1% tail, but of course, other percentages might be of interest. The second type of event, instead of focussing on a fixed

---

[9] Crnkovic and Drachman (1995) note that their proposed K_m statistic can be tailored to the interests of the forecast evaluator by introducing the appropriate weighting function.

percentile 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 q% of y_t over a one-day period. The corresponding probability forecast generated from model m is

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

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

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.[10] 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). A key property of the QPS is that it is a proper scoring rule, which means that forecasters must report their actual forecasts to minimize their expected QPS score. Thus, accurate VaR models are expected to generate lower QPS scores than inaccurate models.

In addition to being intuitively simple, QPS is a useful scoring rule because it highlights

---

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

the three main attributes of probability forecasts: accuracy, calibration and resolution. The QPS can be decomposed as $QPS = QPS_{\bar{R}} + LSB - RES$, where $QPS_{\bar{R}}$ is QPS evaluated with P_t equal to the observed frequency of occurrence for all t. Accuracy refers to the closeness, on average, of the predicted probabilities to the observed realizations and is directly measured by QPS. Calibration refers to the degree of equivalence between the forecasted and observed frequencies of occurrence and is measured by LSB. Resolution is the degree of correspondence between the average of subsets of the probability forecasts with the average of all the forecasts and is measured by RES.

The QPS measure is 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 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 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 methodologies to gauge the accuracy of alternative VaR models and 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 methodology, its ability to correctly classify VaR models is gauged by how frequently the QPS value for the true data generating process is lower than that of the alternative models.

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. For the purposes of this simulation exercise however, the portfolio in question has been simplified.

The simulated portfolio y_t will be a simple integrated process of order one; that is,

$y_t = y_{t-1} + \varepsilon_t$, where ε_t has distribution f_t.

The simulation experiment is conducted in four distinct, yet interrelated, segments. In the first two segments, the emphasis is on the shape of the f_t distribution. To examine how well the various evaluation methodologies perform in the face of 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 methodologies in the presence of variance dynamics in ε_t. Specifically, the third segment uses innovations from a GARCH(1,1)-normal model, and the fourth segment uses innovations from a GARCH(l,1)-t(6) model.

In each segment, the true data generating process is one of 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. For example, a popular type of VaR model specifies its variance h_{mt} as an exponentially weighted moving average of squared innovation; that is,

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

This VaR model, as used in the well-known Riskmetrics calculations (see Guldimann, 1994), is calibrated here by setting λ equal to 0.97 or 0.99. A description of the alternative models used in each segment of the simulation exercise follows.

For the first segment, the true data generating process for ε_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 data generating process is a t(6) distribution. The six alternative models are two normal distributions with variances of 1 and 1.5 (the same variance as the true model), 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. 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., standard normal or 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 similarly. For each run, the simulated y_t series is generated using the chosen data generating process. The chosen length of the in-sample series (after 1000 start-up observations) is 2500 observations, which roughly corresponds to ten years of daily observations. The seven alternative VaR models are then used to generate the necessary 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 forecasts from the various VaR models are then evaluated using the appropriate evaluation methodology. For the binomial and interval forecast methodologies, the four coverage probabilities examined are $\alpha = [1, 5, 10, 25]$. For the distribution forecast methodology, only one null hypothesis can be specified. For the probability forecast methodology, two types of regulatory events are examine. First, using the empirical distribution of ε_t based on the 2500 in-sample observations, 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.[11] 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(α,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 l% 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,

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

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 methodologies are presented for each data generating process and its alternative VaR models. The results are based on a minimum of 1000 simulations.

Three general points can be made regarding the simulation results. First, the power of the three statistical methodologies varies considerably; i.e., 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 methodologies are likely to misclassify inaccurate models as accurate.

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

---

[11] The determination of this empirical quantile of interest is related to, but distinct from, the "historical simulation" approach to VaR model evaluation.

alternative model, the loss function score 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 methodology seem to be low. 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.

Third, for the cases examined, all four evaluation methodologies seem to be more sensitive to misspecifications of the distributional shape of f_t than to misspecifications of the variance dynamics. 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 methodologies 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. However, the low power of the K test against these alternatives may undermine this conjecture.
- 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 and not just in the tail.

Table 2, Panel B contains the five sets of comparative accuracy results for the probability forecast evaluation methodology. 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 methodology is clearly capable of avoiding the misclassification of inaccurate models.

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

Table 3, Panel A presents the power analysis of the three statistical evaluation methodologies 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 exhibit good to moderate results for the N(0, 1) model, but poor results for the N(0, 1.5) model. With respect to the heteroskedastic models, 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 methodology. Overall, the results indicate that this methodology can correctly gauge 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 more clearly 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 q.

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

Table 4, Panel A presents the power analysis of the statistical evaluation methodologies 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 methodologies 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 methodologies 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 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 methodology. Overall, the results indicate that this methodology 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 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 methodologies 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 that mispecifies both the variance dynamics and f_t is easily 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 models with inaccurate distributional forecasts 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 methodology. Once again, the results indicate that the methodology 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 more due 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.

- With respect to the homoskedastic alternatives, the method is able to classify the alternative models a very high percentage of the time; thus, indicating that incorrect modeling of the variance dynamics is well accounted for in this methodology.
- 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 methodologies 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 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 methodology, based on the probability forecast framework discussed by Lopez (1995), was proposed and examined. By avoiding hypothesis testing and instead relying on standard forecast evaluation tools, this methodology 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 methodology can distinguish between VaR models; that is, the probability forecasting methodology seems to be less prone to model misclassification. 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.

Christoffersen, P. F., 1995. "Evaluating Interval Forecasts," Manuscript, Department of Economics, University of Pennsylvania.

Crnkovic, C. and Drachman, J., 1995. "A Universal Tool to Discriminate Among Risk Measurement Techniques," *Risk,* forthcoming.

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

Diebold, F.X. and Lopez, J. A., 1996. "Forecast Evaluation and Combination," Technical Working Paper #192, National Bureau of Economic Research.

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

Guldimann, T., 1994. *RiskMetrics Technical Document,* Second Edition. New York: JP Morgan.

Hendricks, D., 1995. "Evaluation of Value-at-Risk Models Using Historical Data," Manuscript, Federal Reserve Bank of New York.

Kupiec, P., 1995. "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives,* forthcoming.

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.

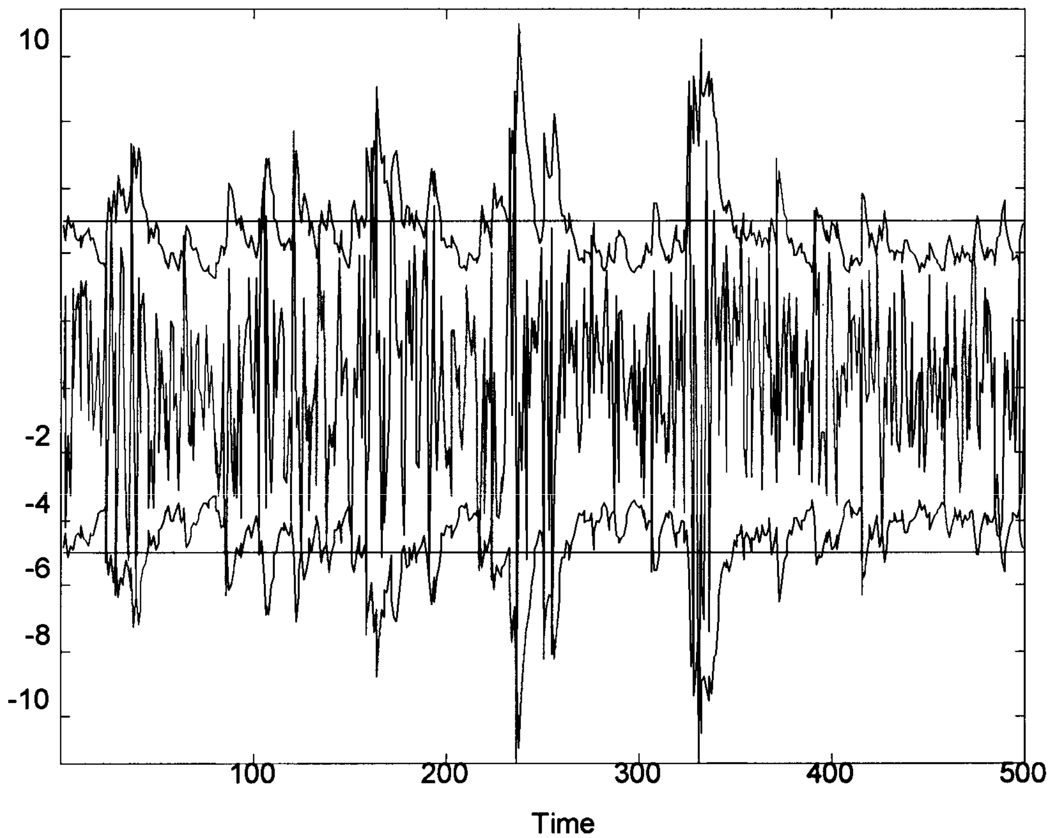
Kupiec, P. And O'Brien, J.M., 1995c. "Recent Developments in Bank Capital Regulation of Market Risks," FEDS Working Paper #95-51, Federal Reserve Board of Governors.

Lopez, J.A., 1995. "Evaluating the Predictive Accuracy of Volatility Models," Research Paper #9524, Research and Market Analysis Group, Federal Reserve Bank of New York.

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

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

Figure 1
GARCH(l, 1) Realization with One-Step-Ahead
90% Conditional and Unconditional Confidence Intervals



This figure graphs a realization of length 500 of a GARCH(l, 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 percentiles 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 Forecasting Methodology[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 a minimum of 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 $\lambda = 0.97$ and $\lambda = 0.99$, 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	67.2	52.1	53.3	62.0	97.5	99.0
Panel B. Accuracy of VaR Models Using the Probability Forecasting Methodology[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 a minimum of 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 Forecasting Methodology[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 a minimum of 1000 simulations. Model 1 is the true data generating process, GARCH(l, 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(l, 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 Forecasting Methodology[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 a minimum of 1000 simulations. Model 1 is the true data generating process, GARCH(l, 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(l, l)-normal model with the same parameter values as Model 1.