Evaluating Value at Risk during the global financial crisis

by

Rahil Valiani

An essay submitted to the Department of Economics

in partial fulfillment of the requirements for

the degree of Master of Arts

Queen's University

Kingston, Ontario, Canada

Abstract

Value at Risk has become one of the most popular measures used by financial institutions to quantify market risk. The goal of this paper is to determine which volatility models are able to accurately estimate VaR during periods of high volatility such as the recent financial crisis. Specifically, we apply three parametric approaches (RiskMetrics, GARCH, EGARCH) under two distributional assumptions (normal, standardized t), and one nonparametric approach (historical simulation) to estimate VaR at 95% and 99% confidence levels. We evaluate the performance of each methodology based on the results of both unconditional and conditional backtesting for the periods before and during the financial crisis. Our results indicate that GARCH and EGARCH with standardized t-distributed residuals are the best models for estimating VaR during the crisis from those tested in this paper.

Keywords: Value at Risk, risk management, financial crisis, volatility models

Acknowledgments

I would like to thank my advisor Wulin Suo for the helpful comments and continuous support throughout this project. I would also like to thank my family for always encouraging me and Ahad Lakhani for his feedback. Lastly, I would like to thank the faculty, staff, and my wonderful classmates at Queen’s over the last five years for a memorable experience.

Table of Contents

1 Introduction:

The past five to seven years have been characterized by significant instabilities in global financial markets. What first began as a subprime mortgage crisis in the United States developed into a global problem affecting numerous countries and financial institutions. The effects of the crisis were seen in equity markets, characterized by dramatic declines in asset and stock prices worldwide; this started in the United States where the S&P 500 dropped 45% between 2007 and 2008 (Bartram & Bodnar, 2009). Amid the widespread effects of the credit crisis, risk management systems in financial institutions came under intense scrutiny, resulting in demands for improvements to their systems.

A widely used risk measure amongst financial institutions is Value at Risk (VaR). VaR has emerged as the industry standard in measuring market risks as well as being used for credit risk, operational risk and liquidity risk (Jorion, 2002). Specifically, VaR is defined as the maximum loss that will be incurred on the portfolio at a given confidence level over a specified period. The measure has gained widespread popularity since 1994 when JP Morgan published RiskMetrics: a system which allows market participants to estimate their exposure to market risk using the "Value-at-Risk framework" through a publically available database online (RiskMetrics technical document, 1994). In 1996, the Basel Committee on Banking Supervision formally introduced a Market Risk Amendment to the Basel Capital Accord which forced banks to use VaR measurements to determine capital requirements (Basel Committee, 1996). Since then, Value at Risk has become an important tool for regulators and has been used to implement policies aimed at preventing severe systemic crises. The popularity of this method is in part due to its simplicity, whereby the risk associated with any portfolio is reduced to just one number that indicates the loss associated to a given probability.

The rise of Value at Risk as the leading market risk measure has led to numerous methodologies being created by academics and practitioners in regards to estimating the measure. These approaches can be summarized into three categories: parametric, nonparametric, and semiparametric. The goal of this paper is to determine which parametric and nonparametric models are able to accurately estimate VaR during periods of high volatility such as the recent financial crisis. Specifically, we apply three parametric approaches (RiskMetrics, GARCH, EGARCH) and one non parametric approach (historical simulation) to estimate VaR at VaR at 95% and 99% confidence levels for six assets: three diversified index portfolios (S&P 500, FTSE 100, Nikkei 225), the spot price of gold, 10-year US Treasury Bills, and a diversified portfolio measuring the performance of the US real estate sector. We subsequently backtest each model using both conditional and unconditional coverage tests to evaluate the accuracy of VaR estimates for the period before and during the financial crisis.

This paper is organized in the following manner. Section 2 provides a thorough introduction to VaR. Sections 3 and 4 present the parametric and nonparametric methodologies used. Section 4 describes the backtesting methodologies employed. Section 5 introduces the data provides summary statistics. Section 6 involves model selection for GARCH and EGARCH volatility models. Section 7 reports backtesting results. Section 8 outlines criticisms of VaR. Section 9 is the conclusion.

2 Value at Risk

We previously state that VaR is defined as the maximum loss that will be incurred on the portfolio at a given confidence level over a specified period. In other words, for a given time horizon t, and confidence level α, the VaR is the loss in market value over the time t that is exceeded with probability $1 - \alpha$. For example, a one day 99% VaR of $1,000,000 means that we expect on 99% of trading days, portfolio losses will not exceed $1,000,000.

While Value at Risk can be used by any entity to measure its risk exposure, commercial and investment banks often use it to capture the potential loss in value of their traded portfolios from adverse market movements. The obtained VaR estimate can then be compared to the banks available capital and cash reserves to ensure that the losses can be covered without putting the firm at risk.

The selection of the confidence level and time horizon plays an essential role in determining VaR and depends on the objectives of the firm. From a regulatory standpoint the Basel Accord requires firms to compute a 99% VaR each trading day. However, depending on the management's risk aversion level, a firm may choose to compute VaR at a different confidence level when setting capital reserves (Jorion, 2001). More risk averse firms set aside a larger capital buffer, which parallels choosing a higher confidence interval. The time horizon that is chosen is dependent on the type of firm being analyzed and the time it takes for this firm to liquidate its portfolio holdings (Jorion, 2001). Banks, for example, have actively traded liquid portfolios and therefore compute VaR daily. Non-financial firms and institutional investors have a longer investment horizon and therefore trade less liquid assets that require long-term loss estimates. From a statistical point of view, VaR can be defined as:

$$VaR_\alpha = \mu + F^{-1}(\alpha) \times \sigma_t, \qquad (1)$$
$$\text{where } \Pr(r_t < -VaR_\alpha) = 1 - \alpha$$

where $F(*)$ denotes the cumulative density function (CDF) which describes the profit and loss distribution of the financial position, μ is the mean return, r_t is the observed return on day t, and σ_t is the standard deviation or volatility. For a T-day VaR, the calculation above is multiplied by the square root of the time horizon T; however, we compute daily VaR and therefore, the time factor is excluded. In order to quantify market risk in terms of monetary amount, VaR is multiplied by the current price of the asset or portfolio.

Since the inception of RiskMetrics, an increasing number of approaches to estimate VaR have been created. Engle and Mangelli (2001) classify the approaches into three categories:

Table 1: Value at Risk Classification

Parametric	**Nonparametric**	**Semiparametric**
▪ RiskMetrics/EWMA ▪ ARCH/GARCH Models	▪ Historical Simulation ▪ Hybrid Model	▪ Extreme Value Theory ▪ CAViaR ▪ Quasi-Maximum Likelihood

Parametric approaches, such as RiskMetrics and GARCH, fit a probability distribution (normal or non-normal) to the data in order to estimate the conditional standard deviation. The standard deviation is then substituted into equation (1) in order to calculate VaR. In comparison, nonparametric approaches directly use the empirical distribution of returns and compute VaR by finding the quantile at a desired confidence level. Section three will discuss both approaches in detail.

Each methodology mentioned above attempts to capture some or all of the characteristics commonly observed in financial data. These empirical regularities of financial data can be summarized as follows (Engle, 2001):

1. Financial return distributions are leptokurtotic, which means they have heavier tails and a higher peak than the normal distribution.
2. Volatilities of market factors tends to cluster, meaning that a period of high price volatility is typically followed by another period of high volatility; while a period of low price volatility is followed by a period of low volatility.[1]
3. Equity returns are typically negatively skewed, implying that observations occur more frequently below the mean than above.

Although there is no industry consensus on the best method for calculating VaR (Engle and Gizycki, 1999), Perignon and Smith (2006) surveyed VaR estimation methods of 60 large banks worldwide over 1996-2006 in order to find which approaches were popular among financial institutions. They found that 73% of banks used the historical simulation methodology while 14% used the Monte Carlo simulation approach. Beder (1995) stated that the results from VaR estimations on a portfolio may vary dramatically depending on the method chosen after applying eight VaR approaches to three portfolios. A strong understanding of the statistical models being employed and their underlying assumptions is therefore required when selecting an appropriate methodology. We compare and contrast various approaches to estimating Value at Risk in the following section.

---

[1] This phenomenon was first noted by Mandelbrot (1963).

3 Parametric Models

As mentioned earlier, financial market data such as the returns of share prices, stock indices and foreign exchange rates have all often been observed to exhibit volatility clustering. In such circumstances, the assumption of constant variance, homoskedasticity, is inappropriate. Moreover, financial time series data often exhibit leptokurtosis, which means that the distribution of their returns is fat-tailed (i.e. relative high probability for extreme values). In this section, we introduce parametric approaches that can be utilized to estimate Value at Risk and deal with the aforementioned phenomena. Each model has its own set of assumptions, along with strengths and weaknesses that affect its ability to estimate VaR. Note that every parametric approach in this paper specifies the following representation of the return series:

$$r_t = \mu + \epsilon_t \tag{2}$$

where μ is the mean return observed and ϵ_t is an unpredictable component which is defined below.

3.1 ARCH Model

Engle (1982) proposed the Autoregressive Conditional Heteroskedasticity (ARCH) model which attempts to capture the volatility clustering feature of financial returns by modeling the dynamics of volatility. ARCH models assume that the current value of the variance depends upon past squared error terms. The model defines an error term or residual, ϵ_t, which can be expressed as an arch process of the following form:

$$\epsilon_t = \sigma_t z_t, \qquad \text{where } z_t \sim i.i.d.(0,1) \tag{3}$$

where $\{z_t\}$, known as the standardized residual, is a white noise process. The conditional variance of ϵ_t is σ_t, a time-varying function of the information set at time $t-1$ (Becketti, 2013). As mentioned earlier, the conditional variance is expressed as a linear function of past squared error terms:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^{\,2} \tag{4}$$

Above is an ARCH (q) process where σ_t^2, the conditional variance, depends on q lags of the squared error term ϵ_t. The unconditional variance, α_0, is a stochastic process and hence remains independent of time. In order to satisfy the non-negativity constraints for the conditional variance, $\alpha_0 > 0$ and $\alpha_i \geq 0$ for $i = 1, \dots, q$.

3.2 GARCH Model

The creation of Engle's model inspired numerous academics and practitioners to construct innovative variations of the original modelling approach. A drawback of the ARCH methodology is that a high order parameterization is usually required in order to capture the dynamics of the conditional variance (Bollerslev, 1986). Many parameters have to be estimated which in addition to being burdensome, presents the possibility of the non-negativity constraint being violated. Bollerslev's (1986) Generalised ARCH (GARCH) model is a natural solution to the problem of high ARCH order parameterizations. The model is based on an infinite ARCH specification and can reduce the number of estimated parameters from an infinite number to just a few to obtain greater parsimony (Becketti, 2013). A GARCH process models conditional variance as a linear function of past squared error terms and lagged conditional variances. The GARCH model can be characterized by the following equation:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^{\,2} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^{\,2} \tag{5}$$

where $\omega > 0, \alpha_i \geq 0$ for $i = 1, \dots, q$, and $\beta_j \geq 0$ for $j = 1, \dots, p$. Described above is a GARCH (p,q) process where σ_t^2, the conditional variance, in addition to depending on q lags of the squared error term ϵ_t, depends on p lags of the conditional variance. To ensure the unpredictable component, ϵ_t, is stationary and conditional variance is positive, the restriction $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$ must hold. The sum of parameters α and β measures the persistence in conditional volatility. For instance, a relatively large value implies that volatility takes a long time to die out following a large shock and revert back towards ω, the long run variance (Jorion, 2001).

3.3 RiskMetrics

The RiskMetrics approach measures the volatility by using an exponentially weighted moving average (EWMA). EWMA assigns a decay factor to volatility in order to place greater emphasis on more recent observations in the time series. Thus, weights for each past return diminish exponentially over time. Exponentially weighted models are able to immediately capture the effects of large fluctuations in the markets. The RiskMetrics model is a special case of an Integrated GARCH (IGARCH), a model which takes into account the existence of a unit root in the variance (non-stationary variance). The IGARCH(1,1) is defined as follows:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)\epsilon_{t-1}^2 \qquad (6)$$

JP Morgan's RiskMetrics methodology sets the exponential decay factor as $\lambda = 0.94$ or $\lambda = 0.97$ for daily and monthly holding periods respectively. Lastly, note that this approach assumes that residuals are normally distributed.

3.4 EGARCH Model

Although GARCH models successfully capture thick tail returns and volatility clustering, they are poor models if we want to capture the leverage effect. The conditional variance is a function of only the magnitudes and not the sign of ϵ_t. The leverage effect refers to the tendency for market declines to forecast higher volatility than equivalent market increases (Black, 1976). Many new models attempt to capture asymmetries in volatility forecasts. These models allow for an asymmetric effect on innovations or unanticipated changes. A popular model used to capture these effects is the EGARCH model proposed by Nelson (1991). As opposed to GARCH models, no restrictions need to be imposed on model estimation since the EGARCH model applies a logarithmic transformation to the variance thereby guaranteeing non-negative forecasts.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q}\left(\alpha_i\left|\frac{\varepsilon_{t-i}}{\sigma_{t-i}}\right| + \gamma_i\frac{\varepsilon_{t-i}}{\sigma_{t-i}}\right) + \sum_{j=1}^{p}(\beta_j \ln(\sigma_{t-j}^{2})) \qquad (7)$$

This is an EGARCH (p,q) process where the parameter γ allows for the asymmetric effect. If $\gamma_1 = 0$, the model is symmetric and good news, $\epsilon_t > 0$, generates the same effect on volatility as bad news, $\epsilon_t < 0$. The presence of the leverage effect can then be explored by testing the hypothesis that $\gamma_1 < 0$. This is representative of a situation where positive shocks affect volatility less than negative shocks**.** Thus, $\gamma_1 > 0$ refers to an opposite scenario where positive shocks affect volatility more than negative shocks.

3.5 Distributional Assumptions

The choice of probability distribution used to model the residuals can have a significant impact on parameter estimation for the models above. In this paper, we apply two distributional assumptions - The standard normal distribution and the standardized t-distribution.

Much of the empirical literature applies the assumption of normally distributed residuals due to its ease of implementation. Although GARCH-type models with normally distributed errors take volatility clustering into account, they do not adequately account for the other stylized facts observed in financial data such as leptokurtosis and negative skewness (Wang and Fawson, 2001). When we calculate VaR, the normal probability density function at the specified confidence level is substituted into equation (1):

$$f(z_t) = \frac{1}{\sigma_t\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_t^2} z_t^2\right) \tag{8}$$

Bollerslev (1987) deviated from the traditional approach of normally distributed ARCH residuals, and proposed the use of the standardized t-distribution. Under the standardized t-distribution, the probability density function used to compute VaR is described by the equation:

$$f(z_t, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi(v-2)}} \left(1 + \frac{z_t^2}{v-2}\right)^{-\frac{v+1}{2}} \tag{9}$$

where v refers to the degrees of freedom parameter. Unlike the normal distribution, the t-distribution allows us to model thicker tails in the distribution and therefore better represents the actual distribution of returns.[2] The degrees of freedom parameter, v, explains the kurtosis of the distribution and is estimated as a parameter when conditional variance is estimated. If $v > 4$, the distribution is considered leptokurtotic; however as $1/v \to 0$, the standardized t-distribution converges to a standard normal distribution. Note that the t-distribution assumption is not without weakness: The distribution is symmetric and therefore unable to model skewness of observed returns.

Although we only look at two distributional assumptions, far more are used in academic literature. Lambert and Laurent (2001) use a skewed t-distribution that takes into account the asymmetry of financial returns unlike the standardized t-distribution. Hull and White (1998) propose a model for computing Value at Risk where the user is permitted to choose any probability distribution for daily returns as long as transformations of the probability density function are multivariate normal. The problem in applying complex distributional assumptions is that estimating inputs is extremely difficult. Furthermore, evaluating the distribution at a given quantile becomes much harder.

4 Nonparametric Models

4.1 Historical Simulation

Unlike parametric volatility models, nonparametric approaches do not assume that asset returns follow a specific probability distribution function. The historical simulation (HS) procedure assumes that the empirical distribution of the past *m* observations ($\{R_{t+1-\varepsilon}\}_{\epsilon=1}^{m}$) approximates the distribution of tomorrow's returns. The $\alpha\%$ Value at Risk is then the αth percentile of the sequence of previous *m* days' returns. The formula for historical simulation VaR is described by the equation below:

$$VaR_{t+1}^{\alpha} = -Percentile\{\{R_{t+1-\epsilon}\}_{\epsilon=1}^{m}, \alpha\} \quad (10)$$

Equivalently, if the returns are sorted in ascending order, VaR_{t+1}^{α} is simply the number such that $\alpha\%$ of the returns are smaller than VaR_{t+1}^{α}. If VaR falls in between two returns, a form of interpolation is applied. We specifically apply linear interpolation to determine the precise value of the empirical

[2] This is due to taking z_t to a power versus an exponential (Bollerslev, 1987).

distribution at the desired quantile. In order to estimate VaR for the following day, the *m* period window is moved forward by one observation and the procedure is repeated.

The main advantage and reason for the approach's popularity is due to its simplicity and ease of implementation (Dowd, 1998). Unlike GARCH-type models, no parameters need to be estimated and therefore no numerical optimization (such as maximum likelihood estimation) is necessary. This is advantageous as the technique does not suffer from the flaw of parametric models where relying on modeling assumptions can be misleading if the model is poorly specified. Although historical simulation does not make an explicit assumption on the distribution of asset returns, the approach makes the implicit assumption that the distribution of returns is the same within the window, and thus remains the same in the future as it has been in the past (Engle, 2001). However, Dowd (1998) argues that this feature makes the approach less restrictive since we don't have to assume that returns are independent. Therefore, the approach is able to accommodate the fat tails inherent in financial data.

The critical limitation of historical simulations lies in its failure to account for the lag effect of volatility by implicitly giving equal weight to each observation in the sample. Another drawback is the choice of *m*, the length of the observation window (Jorion, 2001). If *m* is too large, then recent observations (which are presumably more relevant for tomorrow's distribution) will carry little weight. This is especially dangerous in a crisis period. When the market moves from a period of low volatility to one with high volatility, VaR estimates will be underestimated since it will take some time before the observations from the low volatility period exit the window. On the other hand, if *m* is too small, there may not be enough observations in the left tail to precisely calculate VaR at a high confidence level. Therefore, the selection of *m* plays a major role in the magnitude of VaR when implementing historical simulation. Lastly, keep in mind that for a newly issued asset, the lack of historical data available will restrict the possible window size, possibly rendering the approach useless.

4.2 Hybrid Approach

In the previous section, a major drawback mentioned in regards to the historical simulation approach was that each observation in the window was equally weighted. The hybrid approach proposed by Boudoukh et al. (1998) is a variation of historical simulation which assigns exponentially declining weights to each observation in the window. This is also seen in the RiskMetrics methodology. Once weights have been applied to each observation, they are sorted in ascending order. The α% VaR is estimated by summing the weights of the ascending returns until the α% level is reached. Similar to historical simulation, if VaR falls in between two returns, interpolation is utilized. The hybrid approach doesn't require any parameters to be estimated and has the added advantage of capturing the conditional volatility by applying more weight to recent returns.

5 Backtesting

The models described in the previous section each had their own shortcomings. In order to determine the efficiency and accuracy of Value at Risk forecasts we employ backtesting: A method that compares VaR estimated by a model with the actual profit and loss of an asset or portfolio across the sample of interest. The goal of unconditional backtesting is to check whether the proportion of times VaR is exceeded corresponds to the confidence level. For instance, we would expect a 95% VaR to be exceeded approximately 5% of the time. The issue is that since the number of exceptions or violations (the number of times VaR is exceeded by returns) will not exactly be in line with the confidence level, we must determine a range of exceptions over which we would accept or reject the model. The decision is assessed by weighing the costs between rejecting a correct model (type 1 error) or accepting an incorrect model (type 2 error) (Jorion, 2001).[3] A good model not only provides the correct number of exceptions but also ensures that exceptions do not cluster and are evenly spread out (Christoffersen, 1998). This is checked using conditional coverage tests that examine the conditionality of exceptions. In this paper, we employ three backtesting methodologies: Kupiec's test

[3] Ideally, we prefer a test that has high power, or low probability of accepting an incorrect model (type 2 error).

for unconditional coverage, Christoffersen's test for independence, and Christoffersen's joint test for conditional coverage. We also outline the Basel backtesting framework that banks must adhere to.

5.1 Unconditional Coverage

Kupiec (1995) proposed a test which inspects whether the observed frequency of exceptions is statistically equal to the expected frequency of violations. In order to set up the test, we first define I_t as an indicator function for whether a violation took place as follows:

$$I_t = \begin{cases} 1, & \text{if } R_t < VaR_t \\ 0, & \text{if } R_t \geq VaR_t \end{cases} \tag{11}$$

where I_t is distributed as an $i.i.d.$ Bernoulli random variable. Letting $\{I_t\}_{t=1}^{T}$ be the sequence of violations across the *T* day sample being backtested, we can define $N = \sum_{t=1}^{T} I_t$ as the number of violations over a *T* day period. Thus *N* follows a binomial distribution with parameters T and p, where $p = N/T$ is the expected frequency of violations. A proper model should be unbiased and hence the number of violations should converge to p as $T \rightarrow \infty$ (Kupiec, 1995).

Mathematically, the Kupiec test describes unconditional coverage by testing the null hypothesis $E[I_t] = p$ against the alternative hypothesis $E[I_t] \neq p$ using the likelihood ratio test below (Christoffersen, 1998):

$$LR_{UC} = -2\ln[(1-p)^{T-N}p^{N}] + 2\ln[(1-N/T)^{T} - (N/T)^{N} \sim \chi^2(1) \tag{12}$$

The p-value is then equal to $1 - F(LR_{UC})$. A risk model is considered inadequate if the p-value is below the desired level of significance.

5.2 Conditional Coverage

The main critique of the Kupiec test is that it fails to take into account the conditionality of exceptions. Observed exceptions may cluster which is a major source of concern; for example, if most of the violations occurred in the previous two weeks, current risk levels would be much higher than if the violations were randomly scattered throughout the sample. An unconditional test would be unable to capture the increased volatility and consequently fail to accurately describe the risks being faced. A risk manager should realize that in this situation, the probability of a violation tomorrow is greater than α%. Therefore, models that show the violations are clustered should be rejected.

Christoffersen (1998) developed a joint test for conditional coverage which simultaneously tests for unconditional coverage and whether the violations are serially independent. In order to evaluate conditional coverage, we first describe Christoffersen's test for independence among exceptions below:

$$LR_{IND} = -2\ln[(1-\pi)^{T_{00}+T_{01}}\pi^{T_{01}+T_{11}}] + 2\ln[(1-\pi_0)^{T_{00}}\pi_0^{T_{01}}(1-\pi_1)^{T_{10}}\pi_1^{T_{11}} \sim \chi^2(1) \quad (13)$$

where T_{ij} represents the number of days where state j was observed one day while state i was observed the previous day, and π_i characterizes the probability of observing an exception conditional on state i. The likelihood ratio rejects models that create too many or too few clustered VaR violations. Keep in mind that Christoffersen's test only assesses independence on the basis of whether exceptions occur on consecutive days. A superior test for independence would incorporate checking for a certain number of exceptions in a short interval. Therefore if the model is adequately specified, the probability of an violation today should not depend on whether a violation took place the previous trading day; i.e. π_1 should be equal to π_0.

Finally, the conditional coverage test can be computed as the sum of the test statistics for the unconditional coverage and independence tests:

$$LR_{CC} = LR_{UC} + LR_{IND} \sim \chi^2(2) \quad (14)$$

In this paper we apply both conditional and unconditional backtesting methodologies; however, several other methodologies exist. They include the mixed Kupiec test that measures the time between exceptions, and Lopez's test that utilizes a loss function to inspect the magnitude of the violation compared to VaR (Campbell, 2005).[4]

5.3 Regulatory Backtesting

This section discusses the Basel Committee's rules for backtesting.[5] The Internal Models Approach (IMA) highlighted in the Basel Accord computes the Value at Risk of a financial institution by imposing a 99% confidence level over a 10 day trading horizon. The market risk charge for day t is calculated using the following formula (Philippe, 2011):

$$MRC_t^{IMA} = Max\left(k\,\frac{1}{60}\sum_{t=1}^{60} VaR_{t-1}\,, VaR_{t-1}\right) + SRC_t \qquad (15)$$

where k is a multiplicative factor which effectively increases the level of confidence to account for model misspecifications. Due to the multiplicative factor, the market risk charge will usually be a 60 day average over the previous trading day's VaR. SRC_t denotes the specific risk charge, which provides insurance in the form of a capital buffer against idiosyncratic risks, including financial crises.

The Basel Committee's backtesting framework for the Internal Models Approach uses daily backtesting of VaR at the 99% level over the previous year. According to the backtesting outcome, the Basel Committee sorts financial institutions into three zones: green, yellow, and red. The number of exceptions, as stated in the table below, directly determines the multiplicative factor applied and zone a financial institution will be categorized under (Basel Committee, 1996).Therefore, a bank's capital requirement is directly related to the outcome of the backtest. Similar to the unconditional coverage methodology, the backtesting methodology outlined by the Basel Committee fails to account for conditionality of exceptions. In response to the most recent financial crisis, the Basel Committee added a Stressed VaR (SVaR) to the Market Risk Charge which takes into account recent portfolio losses (Basel Committee, 2011). SVaR is defined as the current portfolio loss that

[4] See Campbell (2005) for a thorough review of other backtesting methodologies.

[5] Any commercial bank with trading activity greater than $1 billion or with trading activity greater than 10% of its total assets must hold regulatory capital against their exposures (Basel Committee on Banking Supervision, 1996).

corresponds to a 99% confidence level over a 10 trading day horizon which encompasses a historical 12 month period of unstable financial markets.

$$MRC_t^{IMA} = Max\left(k\,\frac{1}{60}\sum_{t=1}^{60} VaR_{t-1}\,,VaR_{t-1}\right) + Max\left(k_s\,\frac{1}{60}\sum_{i=1}^{60} SVaR_{t-i}, SVaR_{t-1}\right) + SRC_t \qquad (16)$$

Table 2: Basel Penalty Zones

Zone	Number of Exceptions	k
Green	0 to 4	3
Yellow	5	3.4
	6	3.5
	7	3.65
	8	3.75
	9	3.85
Red	10+	4

Notes: Number of Exceptions are counted over a 250 day period.

6 Data Description

The analysis in this paper is based on three diversified stock market indices (S&P 500, FTSE 100, NIKKEI 225), the spot price of gold, 10 Year US Treasury Bills, and the Dow Jones US Real Estate ETF.[6] These assets were chosen based on their relevance to the financial crisis. Their fundamental differences should provide robustness to the evaluation of each model to avoid results dependent on a specific market. US Treasury Bills and gold are considered to be relatively safer assets for investors than equities in times of market uncertainty (Baur & McDermott, 2010). On the other end, the real estate ETF is directly tied to assets that were affected in the subprime crisis.

Each data series was acquired from Yahoo! Finance or Thomson Reuters DataStream for the period 1 January 1990 to 31 December 2008. We conduct empirical investigations for two periods - A pre-crisis period from 1 January 2004 to 31 December 2007 and a crisis period from 1 January 2008 to 31 December 2008 in order to analyze which models accurately estimate Value at Risk during crisis

[6] Dow Jones US Real Estate ETF is a diversified index portfolio which measures the performance of the real estate sector of the US equity market.

periods. Data prior to 2004 is not considered to avoid the effects of the 'dot.com' bubble in the United States.

In order to conduct meaningful analysis on the asset data, we must first make it stationary. The data in Figure 1 is clearly trending over time, implying that the asset price series' are non-stationary. Not only can this be seen from the data itself, it is also apparent in the sample autocorrelation function (ACF). The autocorrelation function slowly decays, which is a characteristic of a long memory process and is very common among many types of financial data (Engle et. al., 1993). In order to formally check if the data is stationary, we use the Augmented Dickey Fuller (ADF) test for unit roots. Our results indicate that the ADF test fails to reject the null hypothesis of a unit root for prices, which indicates that the data is non-stationary.

Converting the asset price series into an asset returns series removes the trend and results in a stationary time series, verified by the ADF test in Table 3. If P_t denotes the price of an asset at time t, asset returns are defined as follows:

$$R_t = \ln(P_t/P_{t-1}) \tag{17}$$

Figure 2 shows the histogram of daily asset returns with an overlying normal distribution. Note that the histogram has fatter and longer tails, especially on the left side of the distribution**.** This suggests that larger losses occur more frequently than the normal distribution indicates. Table 3 also shows results from the Jacques-Bera and Shapiro-Wilks tests for normality. Both tests reject the null hypothesis of normally distributed returns.

Descriptive statistics are reported for each asset in Table 3 for both pre-crisis and crisis periods. Negative mean returns were realized in the crisis period as all assets lost value except for gold which had an approximate return of zero. This is likely due to its reputation as a secure asset in turbulent times. Additionally, assets were far riskier in 2008 during the financial crisis as standard deviations for all six assets at the very least doubled. This is verified by Figure 1 where standard deviation or volatility of each returns series substantially increased in the crisis period.

Table 2: Summary Statistics

	S&P 500		FTSE 100		NIKKEI 225		TREASURY		HOUSING		GOLD	
	Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis
Summary Stats												
Observations	1005	252	1010	252	982	244	1004	252	1006	252	1043	262
Min	-0.0353	-0.0947	-0.0419	-0.0926	-0.0557	-0.1211	-0.0509	-0.0846	-0.0589	-0.2307	-0.0554	-0.0714
Max	0.0288	0.1096	0.0344	0.0938	0.0360	0.1323	0.0597	0.0916	0.0458	0.1511	0.0399	0.0680
Mean	0.0003	-0.0019	0.0004	-0.0015	0.0004	-0.0022	-0.0001	-0.0022	0.0004	-0.0020	0.0007	0.0000
Median	0.0008	0.0000	0.0008	-0.0014	0.0004	-0.0009	0.0000	-0.0025	0.0011	-0.0027	0.0008	0.0001
Std. Dev.	0.0076	0.0259	0.0080	0.0237	0.0111	0.0294	0.0110	0.0254	0.0121	0.0451	0.0107	0.0201
Skewness	-0.3104	-0.0391	-0.4269	0.1265	-0.3619	-0.2320	-0.0466	-0.0898	-0.5268	-0.4699	-0.6114	-0.1411
Kurtosis	4.7961	6.6618	5.7802	6.3230	4.4159	6.6924	4.8435	4.4119	4.6822	6.8079	5.5942	4.3735
Unit root test												
ADF	-75.03	-34.737	-72.686	-35.767	-31.581	-16.632	-30.638	-15.019	-30.111	-19.192	-32.19	-15.839
(p-value)	0	0	0	0	0	0	0	0	0	0	0	0
Normality tests												
Jacques Bera	48.95	22.27	-	21.05	43.63	23.48	38.09	9.45	66.67	29.44	-	9.84
(p-value)	0	0	0	0	0	0	0	0.0089	0	0	0	0.0073
Shapiro-Wilk	6.56	5.981	13.713	7.658	5.589	5.516	5.817	3.205	6.713	5.698	7.952	3.18
(p-value)	0	0	0	0	0	0	0	0	0	0	0	0.00074

Notes: ADF refers to the Augmented Dickey Fuller Test for stationarity. Normal refers to the period between 01/01/2004 to 31/12/2007 while Crisis refers to 01/01/2008 to 31/12/2008.

Skewness indicates whether the data is symmetric or not. Except for the FTSE 100 during the crisis period, each asset is negatively skewed in both subsamples which implies that observations occur more frequently below the mean than above it. These results are consistent with the vast empirical evidence that return distributions of financial assets are asymmetric. Therefore, the normal distribution, which is symmetrically distributed around its mean, is not a realistic assumption for a volatility model and can lead to the underestimation of VaR.

Kurtosis is a measure of the "peakedness" of a distribution and the heaviness of its tails (Wang & Fawson, 2001). As stated previously, financial return distributions generally have high kurtosis and therefore exhibit leptokurtosis. This means that they have fatter tails and a higher peak than the normal distribution. All assets have kurtosis greater than three (kurtosis of normal distribution) in both periods, and therefore, display excess kurtosis. It is worth noting that kurtosis in the crisis period is higher for all assets except for gold and US Treasuries. This implies that larger shifts in asset prices occurred more frequently throughout the crisis than in previous years. We attempt to model some of the leptokurtosis by applying standardized t-distributed residuals when modeling asset returns.

7 Model Specification for GARCH and EGARCH

Before attempting to fit a GARCH or EGARCH model to the data, we first formally test for the presence of ARCH effects using the Lagrange Multiplier test. Whenever the Lagrange Multiplier test gives evidence of ARCH effects for $q > 4$ lags, as it did for each asset, a GARCH model is more appropriate than an ARCH model (Becketti, 2013). Therefore, we skip the fitting of ARCH models entirely. Next, we model the conditional variance as a GARCH(p, q) or EGARCH(p, q) process where $p = 1, 2$ and $q = 1, 2$ and residuals follow either a normal or standardized t-distribution. In total, we estimate 96 models. Model parameters are estimated using the maximum likelihood estimation (MLE) method which maximizes the log-likelihood function. Once the models have been fitted to the data, we select the optimal lag length using Akaike's information criterion (AIC) and Bayesian information criterion (BIC). Davidson and McKinnon (2004) state that whenever two or models are nested, the AIC may fail to choose the most parsimonious one. Thus, the BIC that favours parsimonious specifications may lead to a better fit. We then discard any model with either statistically insignificant coefficients at the 5% level or for which the log-likelihood function failed to converge.

According to these criteria, the $p = 1, q = 1$ specification is optimal for each returns series. Table 4 reports that models with t-distributed residuals return a lower AIC and BIC than the normally distributed alternative. Overall, the remaining models were ranked in terms of their ability to model conditional volatility as follows: EGARCH(1,1) with t-distribution, GARCH(1,1) with t-distribution, EGARCH(1,1) with normal distribution, and GARCH(1,1) with normal distribution. The superiority of the EGARCH model is explained by the asymmetry in the volatility. The leverage effects γ are negative and significant; therefore, negative shocks produced greater volatility than a positive shock of equivalent magnitude.

Next, we evaluate the adequacy of the model using the standardized residuals.[7] If the volatility equation is correctly specified, the squared standardized residuals should not display serial correlation or conditional heteroskedasticity (Becketti, 2013). One method used to test this was the portmanteau test for white noise. The test shows no evidence that the standardized residuals deviate from white noise for each case as we are unable to reject the null hypothesis of no serial correlation. The second test executed for model adequacy was the Lagrange Multiplier test that was run earlier to test for ARCH effects.[8] The results indicate that we are unable to reject the null hypothesis of no ARCH effects, which leads us to conclude that all models adequately model volatility. Note that the aim of this paper is not to select models that best describe conditional volatility of assets, but which accurately estimate VaR following previously described backtesting criteria. Results from backtesting are analyzed in the next section.

[7] Residuals divided by standard deviation.
[8] Test was executed after running a regression on the standardized residuals.

Table 3: GARCH(1,1) and EGARCH(1,1) parameter estimates assuming normal or t-distributed residuals

	S&P 500	FTSE 100	NIKKEI 225	US TREASURY	HOUSING	GOLD
			GARCH(1,1): Normally Distributed Residuals			
μ	0.0005	0.0004	0.0004	-0.0002	0.0009	0.0000
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)
α_1	0.0571	0.0893	0.1063	0.0447	0.1502	0.0484
	(0.0038)	(0.0069)	(0.0061)	(0.0033)	(0.0138)	(0.0021)
β_1	0.9371	0.8995	0.8881	0.9511	0.8450	0.9532
	(0.0043)	(0.0079)	(0.0061)	(0.0037)	(0.0131)	(0.0019)
AIC	-33448	-33439	-28962	-32047	-13290	-27284
BIC	-33422	-33413	-28935	-32021	-13267	-27259
			GARCH(1,1): t-Distributed Residuals			
μ	0.0006	0.0005	0.0003	-0.0003	0.0010	0.0000
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)
α_1	0.0556	0.0832	0.0931	0.0442	0.1494	0.0669
	(0.0060)	(0.0083)	(0.0086)	(0.0053)	(0.0187)	(0.0084)
β_1	0.9431	0.9067	0.9060	0.9548	0.8462	0.9423
	(0.0060)	(0.0092)	(0.0083)	(0.0054)	(0.0176)	(0.0059)
v	6.5940	12.9472	7.7040	6.3445	9.4238	3.8156
	(0.5229)	(1.7695)	(0.7261)	(0.5984)	(1.7174)	(0.2729)
AIC	-33730	-33503	-29138	-32260	-13333	-27789
BIC	-33698	-33470	-29105	-32227	-13305	-27757
			EGARCH(1,1): Normally Distributed Residuals			
μ	0.0002	0.0001	0.0000	-0.0004	0.0007	0.0001
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0001)
α_1	0.1108	0.0445	0.1735	0.0921	0.2579	0.1107
	(0.0085)	(0.0025)	(0.0106)	(0.0066)	(0.0188)	(0.0052)
β_1	0.9826	-0.9321	0.9763	0.9945	0.9833	0.9916
	(0.0017)	(0.0036)	(0.0025)	(0.0011)	(0.0030)	(0.0009)
γ	-0.0909	-0.0066	-0.0970	-0.0290	-0.0589	0.0523
	(0.0060)	(0.0008)	(0.0059)	(0.0034)	(0.0125)	(0.0035)
AIC	-33592	-31571	-29133	-32066	-13294	-27311
BIC	-33560	-31538	-29101	-32033	-13266	-27279
			EGARCH(1,1): t-Distributed Residuals			
μ	0.0004	0.0003	0.0000	-0.0004	0.0008	0.0001
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)
α_1	0.1092	0.1300	0.1617	0.0881	0.2659	0.1306
	(0.0119)	(0.0133)	(0.0145)	(0.0100)	(0.0271)	(0.0138)
β_1	0.9883	0.9870	0.9802	0.9971	0.9836	0.9955
	(0.0021)	(0.0024)	(0.0031)	(0.0014)	(0.0045)	(0.0017)
γ	-0.0884	-0.0721	-0.0956	-0.0319	-0.0566	0.0429
	(0.0085)	(0.0075)	(0.0087)	(0.0057)	(0.0170)	(0.0091)
v	7.3224	13.8852	9.0254	6.4509	9.7652	3.8911
	(0.6031)	(1.9381)	(0.9696)	(0.6014)	(1.8944)	(0.2798)
AIC	-33830	-33580	-29258	-32281	-13332	-27801
BIC	-33791	-33541	-29219	-32242	-13298	-27764

Notes: μ is the constant mean return. α_1 is the ARCH coefficient. β_1 is the GARCH coefficient. v refers to the degrees of freedom parameter for the t-distribution. γ refers to the asymmetric effect of returns. AIC refers to the Akaike Information Criterion and BIC refers to the Bayesian Information Criterion. Standard errors are in parentheses.

8 Backtesting Results

8.1 RiskMetrics

Backtesting results for 95% and 99% VaR computed using the RiskMetrics model are displayed in Table 5. Our results indicate that during the pre-crisis period, we fail to reject the Kupiec test for 95% VaR for all assets. With respect to 99% VaR, we fail to reject the Kupiec test for 4 out of the 6 assets. One explanation for weak results at the 99% confidence level is that the EWMA model assumes normally distributed residuals. As explained previously in the paper, there exists substantial evidence that the returns are not normally distributed. If returns are not normally distributed, the assumption of normal returns will be unable to capture the outliers in the actual return distribution; VaR estimates will be too conservative and understate the actual risk of the asset. As expected, the RiskMetrics model produces worse results during the financial crisis due to the substantial increases in volatility during this period. Across both confidence levels, we reject the Kupiec test for 33% of estimates. Note that the choice of decay factor plays a vital role in the speed at which VaR estimates react to significant changes in volatility. We previously stated that RiskMetrics sets the decay factor, λ, equal to 0.94; but a lower decay factor would incorporate recent volatility faster and may possibly provide better results in the crisis period. The disadvantage, however, is that when a low decay factor is applied, fewer observations carry significant weight. This could result in missing crucial data points that should be included.

In order to examine whether the violations are independent or if they cluster, we apply a likelihood ratio test developed by Christoffersen (1998). For both periods, we fail to reject the null hypothesis of independence between exceptions for 80% of estimates. Although the independence test provides good results, it is important to remember that it only tests for dependence by checking if exceptions occur on consecutive days. A superior test for independence would incorporate checking for a certain number of breaks in a short interval. While violations do not occur on consecutive days, we noticed clustered VaR breaks when manually checking the data, confirming the weakness. Subsequently, we apply Christoffersen's joint test for conditional coverage, which not only examines the exception rate, but the independence of exceptions as well. Moreover, Campbell (2005) states that it is good practice to run the Kupiec and the independence test separately since it is possible that in some cases the model will pass the joint test while failing the others. For the 95% confidence level VaR, we reject conditional coverage for 25% of estimates; while for the 99% confidence level VaR, we reject it for 42% of estimates. Since the conditional coverage is a joint test that sums the likelihood ratios from unconditional coverage and independence tests, the poor results are driven by the weak unconditional coverage during the crisis as found by the Kupiec test.

Table 5: RiskMetrics backtesting results during normal and crisis periods

	Percentage of Violations		Kupiec Test		Independence Test		Christoffersen's Test	
	Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis
				95% Confidence Level				
S&P 500	5.67%	7.94%	0.74	0.05	0.64	0.06	0.85	0.03*
FTSE 100	5.35%	9.92%	0.61	0.08	0.58	0.21	0.76	0.10
NIKKEI 225	4.89%	8.20%	0.13	0.04*	0.72	0.56	0.29	0.09
US TBILL	5.27%	5.95%	0.86	0.34	0.02*	0.08	0.08	0.14
HOUSING	5.57%	7.94%	0.80	0.05*	0.84	0.06	0.95	0.03*
GOLD	5.37%	6.35%	0.80	0.34	0.04*	0.00**	0.12	0.00**
				99% Confidence Level				
S&P 500	2.19%	3.57%	0.37	0.00**	0.56	0.36	0.57	0.00**
FTSE 100	1.29%	3.17%	0.09	0.01**	0.02*	0.47	0.02*	0.02*
NIKKEI 225	1.63%	2.87%	0.50	0.15	0.59	0.65	0.69	0.32
US TBILL	1.89%	1.59%	0.08	0.06	0.47	0.59	0.17	0.15
HOUSING	1.89%	2.38%	0.04*	0.06	0.44	0.59	0.10	0.15
GOLD	2.29%	1.98%	0.01**	0.17	0.06	0.01**	0.00**	0.02*

Notes: Percentage of violations refers to the percentage of observations where losses exceeded VaR. Christoffersen's Test refers to the joint test for conditional coverage. Normal refers to the period between 01/01/2004 to 31/12/2007 while Crisis refers to 01/01/2008 to 31/12/2008. Results from the three tests are expressed as p-values. $^*p<0.05$, $^{**}p<0.01$.

8.2 GARCH Model

Table 6 reports backtesting results for the GARCH(1,1) model with both normally and t-distributed residuals. The assumption of normality produces good 95% VaR estimates prior to the financial crisis; however, exception rates during the crisis are greater than 5% for every asset, with the Nikkei 225 providing violations for 8.20% of observations. Weak results are once again given due to the normality assumption being unable to capture the degree of leptokurtosis in asset returns witnessed during the crisis. The model's greatest shortcoming is in its inability to accurately estimate VaR at the 99% confidence level. In the crisis period specifically, gold produces the best unconditional coverage at 1.98% - almost double the percentage of exceptions expected under the Kupiec test. One feasible explanation for why the model produces adequate 95% VaR estimates but inaccurate 99% VaR is that there is a hump in the tail of the distribution of actual returns. This suggests that although a normal distribution produces better results at the 95% confidence level, it underestimates risk at a high confidence level of 99%.

We now turn our attention to the model with standardized t-distributed residuals. For both pre-crisis and crisis subsamples, VaR estimates based on the assumption of t-distributed residuals outperform those derived from the normal distribution at both the 95% and 99% confidence level. This result is not surprising since the t-distribution is better equipped to model tail thickness of the observed returns. Our results indicate that the percentage of violations decreases in all cases, while the Kupiec test returns higher p-values. The 95% VaR estimate for the Nikkei 225 during the pre-crisis subsample is the only instance where unconditional coverage is rejected. However, the reason for unconditional coverage being rejected is not due to the excess number of VaR exceptions, but because there are too few (percentage of exceptions was 3.56%). This implies that either our model simply returns an abnormally small number of violations or that our model is too conservative and overestimating risk. If it's the latter and our model systematically overestimates risk, a less biased model should be favoured. The problem is that detecting systematic biases is extremely difficult at high confidence levels since exceptions are rare events. (Jorion, 2001). Keep in mind that although overstating risk is not nearly as dangerous as understating risk, it means that the financial institution's Market Risk Charges are unnecessarily high, which can be seen as an inefficient allocation of capital. However, some financial institutions may prefer to report high VaR in order to avoid a large multiplier when regulators compute Market Risk Charges.

Similar to RiskMetrics, we observe strong results for Christoffersen's independence test and joint test for conditional coverage. We find that when residuals are normally distributed, 79% of estimates fail to reject the null hypothesis of independence; while 71% do the same for conditional coverage across both pre-crisis and crisis periods. Under the t-distribution, we fail to reject the null hypothesis for either test in both periods for every asset aside from gold. This is a positive sign that the model adequately accounts for volatility clustering inherent in the data.

Table 7: GARCH(1,1) backtesting results during normal and crisis periods

		Percentage of Violations		Kupiec Test		Independence Test		Christoffersen's Test	
		Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis
					95% Confidence Level				
S&P 500	*N*	4.78%	7.94%	0.74	0.05*	0.64	0.06	0.85	0.03*
	T	3.98%	7.54%	0.12	0.08	0.30	0.08	0.18	0.05*
FTSE 100	*N*	4.65%	7.54%	0.61	0.08	0.58	0.21	0.76	0.10
	T	4.06%	7.14%	0.16	0.14	0.79	0.53	0.35	0.28
NIKKEI 225	*N*	3.97%	8.20%	0.13	0.04*	0.72	0.56	0.29	0.09
	T	3.56%	6.15%	0.03*	0.43	0.52	0.93	0.08	0.73
US TBILL	*N*	4.88%	6.35%	0.86	0.34	0.02*	0.08	0.08	0.14
	T	3.98%	5.56%	0.12	0.69	0.07	0.80	0.06	0.89
HOUSING	*N*	5.17%	7.94%	0.80	0.05*	0.84	0.06	0.95	0.03*
	T	4.08%	5.56%	0.17	0.69	0.80	0.20	0.37	0.41
GOLD	*N*	5.17%	6.35%	0.80	0.34	0.04*	0.00**	0.12	0.00**
	T	4.28%	4.37%	0.28	0.64	0.00**	0.00**	0.00**	0.00**
					99% Confidence Level				
S&P 500	*N*	1.29%	3.97%	0.37	0.00**	0.56	0.36	0.57	0.00**
	T	0.70%	1.59%	0.31	0.39	0.75	0.72	0.56	0.65
FTSE 100	*N*	1.58%	3.17%	0.09	0.01**	0.02*	0.47	0.02*	0.02*
	T	0.99%	2.38%	0.97	0.06	0.65	0.59	0.90	0.15
NIKKEI 225	*N*	1.22%	2.05%	0.50	0.15	0.59	0.65	0.69	0.32
	T	0.61%	2.05%	0.19	0.15	0.79	0.65	0.40	0.32
US TBILL	*N*	1.59%	2.38%	0.08	0.06	0.47	0.59	0.17	0.15
	T	0.70%	1.19%	0.31	0.77	0.75	0.79	0.56	0.92
HOUSING	*N*	1.69%	2.38%	0.04*	0.06	0.44	0.59	0.10*	0.15
	T	1.00%	1.59%	0.99	0.39	0.65	0.72	0.90	0.65
GOLD	*N*	1.99%	1.98%	0.01**	0.17	0.06	0.01**	0.00*	0.02*
	T	1.29%	1.98%	0.37	0.17	0.56	0.01**	0.56	0.02*

Notes: *N* refers to normally distributed residuals. *T* refers to *t*-distributed residuals. Percentage of violations refers to the percentage of observations where losses exceeded VaR. Christoffersen's Test refers to the joint test for conditional coverage. Results from the three tests are expressed as p-values. Normal refers to the period between 01/01/2004 to 31/12/2007 while Crisis refers to 01/01/2008 to 31/12/2008.

8.3 EGARCH Model

The backtesting results for the EGARCH(1,1) are reported in Table 7. In the model specification section, we found that the EGARCH(1,1) returned a negative value for γ that was statistically significant. This means that positive shocks generate less volatility than negative shocks of equal magnitude. Thus, it is astonishing that although our model is able to capture the leverage effect, it performs poorly during the crisis period at both levels of confidence in comparison to the GARCH(1,1) - Normal. During the crisis subsample, we reject 4 out of 6 assets when examining the Kupiec test for 95% VaR. With respect to 99% VaR, we reject the Kupiec test for all assets. The S&P 500 provides horrendous results as we find exception percentages of 10.71% and 3.57% for 95% and 99% VaR respectively. After applying the EGARCH model with t-distributed residuals, our results indicate that the model produces far superior estimates of Value at Risk. The only 2 instances where we reject the Kupiec test are witnessed prior to the financial crisis for 95% VaR - these are due to the overestimation of risk. Overall, EGARCH(1,1) with t-distributed residuals provides slightly better p-values and hence performs marginally better than its t-distributed GARCH(1,1) counterpart.

When testing for independence between violations, we find that the EGARCH returns results similar to the GARCH model. However, results for the joint test for conditional coverage are not encouraging in cases where residuals are normally distributed. Throughout the crisis period, we find low p-values for both confidence levels - the highest of which is 0.14. As previously explained, this is clearly driven by the weak unconditional coverage during the crisis as found by the Kupiec test. Conversely, under the t-distribution, the results from Christoffersen's conditional test imply that we are able to jointly accept unconditional coverage and independent exceptions in most instances. Outside of gold, the only other 2 occurrences where conditional coverage is not attained are when our model rejects the Kupiec test due to overestimating risk. Therefore, those particular failures of the Christoffersen test are not reasons for serious concern.

After comparing and contrasting across all three parametric models discussed above, the key takeaway message to focus on is that the distributional assumption for returns plays a much larger role than the type of volatility model being implemented. Therefore, we suggest the use of GARCH or EGARCH models with t-distributed residuals over RiskMetrics.

Table 8: EGARCH(1,1) backtesting results during normal and crisis periods

		Percentage of Violations		Kupiec Test		Independence Test		Christoffersen's Test	
		Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis
				95% Confidence Level					
S&P 500	*N*	4.88%	10.71%	0.86	0.00**	0.69	0.16	0.91	0.00
	T	4.28%	7.14%	0.28	0.14	0.90	0.10	0.56	0.08
FTSE 100	*N*	5.45%	8.33%	0.52	0.03*	0.00**	0.34	0.00**	0.05
	T	4.16%	7.54%	0.21	0.08	0.84	0.63	0.44	0.20
NIKKEI 225	*N*	3.77%	8.61%	0.06	0.02*	0.62	0.48	0.16	0.05
	T	3.05%	6.56%	0.00**	0.29	0.31	0.96	0.01**	0.56
US TBILL	*N*	4.78%	6.35%	0.74	0.34	0.03*	0.08	0.09	0.14
	T	3.68%	5.56%	0.04*	0.69	0.09	0.21	0.03*	0.42
HOUSING	*N*	5.27%	7.54%	0.69	0.08	0.47	0.08	0.71	0.05
	T	3.98%	5.95%	0.12	0.50	0.75	0.17	0.29	0.31
GOLD	*N*	4.48%	7.94%	0.44	0.05*	0.00**	0.02*	0.01**	0.01
	T	4.18%	4.76%	0.22	0.86	0.00**	0.00**	0.00**	0.00
				99% Confidence Level					
S&P 500	*N*	1.09%	3.57%	0.77	0.00**	0.62	0.41	0.85	0.00**
	T	0.80%	1.19%	0.28	0.14	0.90	0.10	0.56	0.08
FTSE 100	*N*	2.48%	3.17%	0.00**	0.01**	0.00**	0.47	0.00**	0.02*
	T	0.89%	1.98%	0.72	0.17	0.69	0.65	0.87	0.35
NIKKEI 225	*N*	1.12%	2.46%	0.71	0.05*	0.62	0.58	0.82	0.13
	T	0.61%	1.23%	0.19	0.73	0.79	0.78	0.40	0.91
US TBILL	*N*	1.39%	2.78%	0.24	0.02*	0.53	0.53	0.41	0.05*
	T	0.70%	0.79%	0.31	0.73	0.75	0.86	0.56	0.93
HOUSING	*N*	1.69%	3.57%	0.04*	0.00**	0.44	0.41	0.10	0.00**
	T	0.90%	1.98%	0.73	0.17	0.69	0.65	0.87	0.35
GOLD	*N*	1.89%	3.17%	0.01**	0.01**	0.10	0.12	0.01**	0.01**
	T	1.19%	1.98%	0.55	0.17	0.50	0.01**	0.67	0.02*

Notes: N refers to normally distributed residuals. T refers to t-distributed residuals. Percentage of violations refers to the percentage of observations where losses exceeded VaR. Christoffersen's Test refers to the joint test for conditional coverage. Results from the three tests are expressed as p-values. Normal refers to the period between 01/01/2004 to 31/12/2007 while Crisis refers to 01/01/2008 to 31/12/2008.

8.4 Historical Simulation

Table 8 reports backtesting results for the historical simulation approach with a 250 day trading window. We previously explained that the historical simulation methodology does not assume that asset returns follow a specific probability distribution. Instead, it uses the nonparametric empirical distribution function of the previous 250 trading days to determine VaR for a specified confidence level. The danger is that this method implicitly assumes that the distribution of past returns provides a complete representation of expected future returns. Historical simulation also equally weighs each observation in the rolling window. This can be especially dangerous during crisis times when there is a trend of increasing volatility that could possibly lead to a downward biased VaR. For the subsample prior to the financial crisis, both 95% and 99% VaR estimates calculated by the historical simulation produce an accurate number of exceptions according to the Kupiec test. Unsurprisingly, during the crisis period we reject the null hypothesis of the Kupiec test for every asset in nearly every instance. In fact, not only do we reject the test, but we encounter extremely low p-values (mostly below 1%) suggesting that we strongly reject unconditional coverage in the crisis.

As found with the parametric VaR approaches, historical simulation provides good results for Christoffersen's test of independence. Across both periods and confidence intervals, 20 out of 24 estimates fail to reject independent VaR violations. Nonetheless, just as we noticed in other approaches, when unconditional coverage is not satisfied, outcomes from the conditional coverage test tend to be poor. In the crisis period, we reject conditional coverage for both 95% and 99% VaR across all assets except 99% VaR FTSE 100. In conclusion, for times where markets are volatile and extremely risky, historical simulation is a poor approach which fails to adequately capture the importance of recent observations. The methodology's largest flaw is not incorporating the lag effect of volatility.

Table 6: Historical Simulation backtesting results during normal and crisis periods

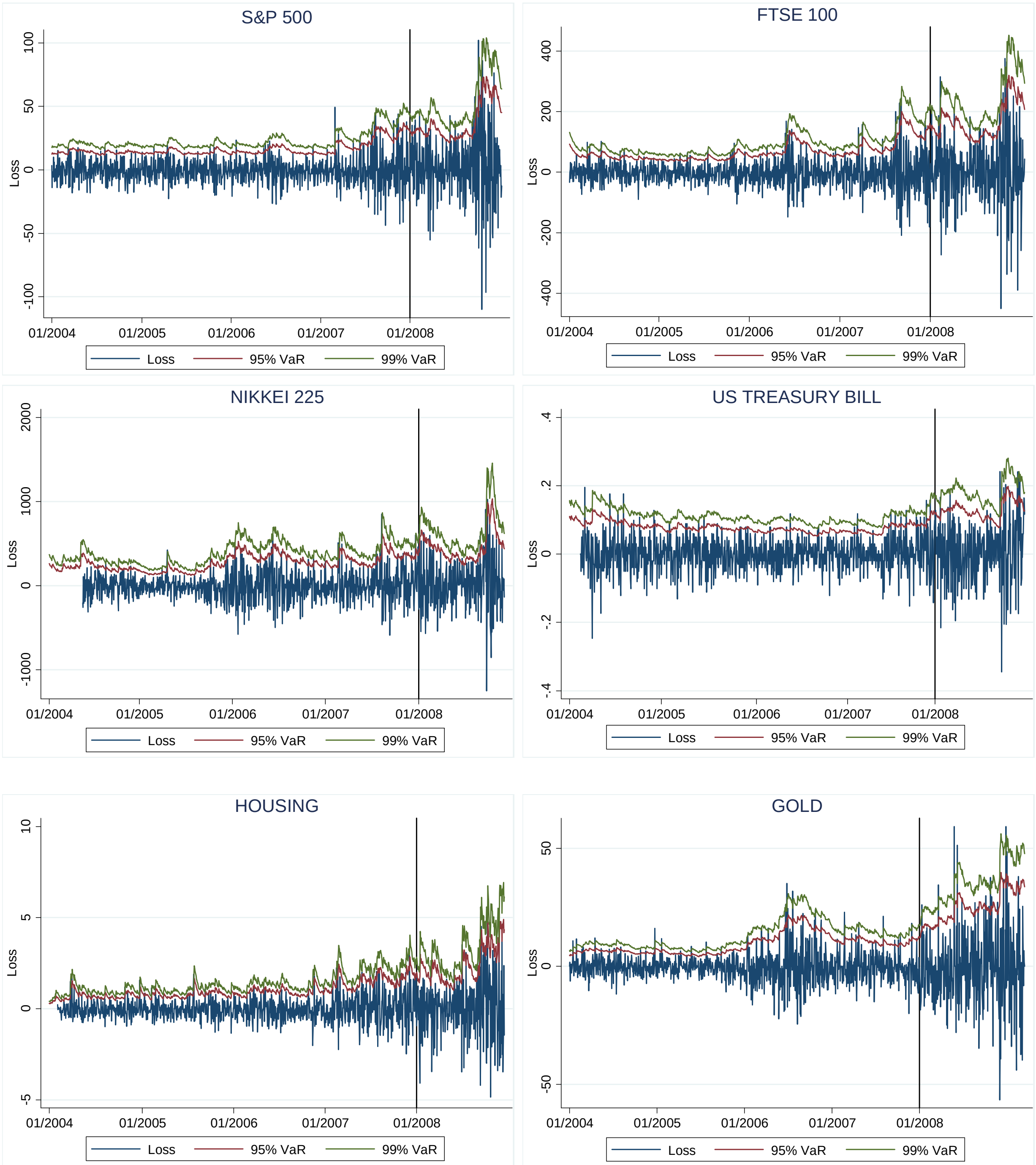
	Percentage of Violations		Kupiec Test		Independence Test		Christoffersen's Test	
	Normal	Crisis	Normal	Crisis	Normal	Crisis	Normal	Crisis
				95% Confidence Level				
S&P 500	5.77%	11.51%	0.27	0.00**	0.06	0.69	0.09	0.00**
FTSE 100	5.35%	9.92%	0.62	0.00**	0.23	0.73	0.43	0.01**
NIKKEI 225	4.48%	11.48%	0.45	0.00**	0.00**	0.42	0.01**	0.00**
US TBILL	5.17%	12.70%	0.80	0.00**	0.22	0.29	0.45	0.00**
HOUSING	6.47%	11.90%	0.04*	0.00**	0.00**	0.80	0.00**	0.00**
GOLD	5.77%	11.90%	0.27	0.00**	0.02*	0.33	0.04*	0.00**
				99% Confidence Level				
S&P 500	1.39%	3.97%	0.24	0.00**	0.53	0.36	0.41	0.00**
FTSE 100	1.29%	2.38%	0.38	0.06	0.56	0.59	0.57	0.15
NIKKEI 225	0.81%	3.69%	0.55	0.00**	0.05*	0.32	0.12	0.00**
US TBILL	1.19%	3.17%	0.55	0.01**	0.59	0.47	0.72	0.02*
HOUSING	1.29%	4.76%	0.37	0.00**	0.56	0.11	0.57	0.00**
GOLD	1.29%	1.98%	0.37	0.17	0.15	0.01**	0.24	0.02*

Notes: Percentage of violations refers to the percentage of observations where losses exceeded VaR. Christoffersen's Test refers to the joint test for conditional coverage. Results from the three tests are expressed as p-values. Normal refers to the period between 01/01/2004 to 31/12/2007 while Crisis refers to 01/01/2008 to 31/12/2008.

7.5 Graphical Results

We state in Section 7.3 that we suggest GARCH or EGARCH models with t-distributed residuals over RiskMetrics. Due to historical simulations ineffectiveness during the crisis, we recommend GARCH or EGARCH models with t-distributed residuals as the best models for estimating VaR from those tested in this paper. Figure 1 shows the profit and loss of each asset along with 95% and 99% Value at Risk using the GARCH(1,1) with t-distributed residuals. Notice the steep rise in volatility at the end of 2007 and throughout 2008 caused by the financial crisis. Not only is the number of VaR exceptions substantially higher in 2008, so too is the presence of clustered violations.

Figure 3: Loss of each asset, along with 95% and 99% VaR



9 Criticisms of VaR:

Although VaR has solidified its place as the industry standard for measuring market risk, it is extremely dangerous to solely rely on this metric when evaluating exposures. We stated earlier that an advantage of Value at Risk is that it aggregates risks across a financial institution into a single number which can be simply understood by those less technically inclined. However, the simplicity of VaR is also one of its biggest weaknesses. A 99% VaR is the monetary amount such that for 1% of observations, loss is expected to exceed VaR. The problem with this statement is that there is no mention about the potential magnitude of losses when VaR is breeched or what the maximum possible loss due to an adverse event could be. For instance, one possible scenario is that the loss is only slightly greater than VaR, while another is that it could wipe out the equity of a company and force it into bankruptcy. There are a number of documented cases of financial institutions being subject to catastrophic events. Famous examples include Barings Bank, and more recently, Lehman Brothers.

One proposed metric to deal with this shortcoming is Expected Shortfall (also known as conditional VaR or CVaR), which is defined as the expected loss given that VaR has been exceeded (Artzner et al., 1997).[9] It focuses on evaluating risk in the left tail of the return distribution and computes the measure in the form of a conditional probability: $E[Loss|Loss < VaR_{\alpha}]$. For example, suppose that in the sample of returns, the average loss of the worst 1% of outcomes is $1,000,000; then our Expected Shortfall or CVaR at the 99% confidence level is $1,000,000. This again relates to a common theme throughout this paper: accurately modelling the distribution of returns; specifically the left tail. Extreme values are important in risk-management as they are associated with catastrophic events such as market crashes. Another popular approach is Extreme Value Theory, a branch of statistics which focuses on modelling the extreme quantiles of a probability distribution. Simply put, this is achieved by modeling the probability distribution of standardized returns in excess of a pre-defined threshold. The main result of EVT is that the extreme tail of a number of distributions can be characterized by the Generalized Pareto distribution (Engle, 2001).[10] Take note that these statistical methodologies utilize historical data, and similar to VaR, cannot predict major shocks.

[9] Basel Committee suggested the possibility of replacing VaR with Expected Shortfall (Basel Committee, 2012).
[10] See Diebold, Schuermann, Stroughair for a thorough review on extreme value theory used in financial risk management.

The recent financial crisis brought much scrutiny over the entire field of risk management, and VaR specifically became the subject of intense criticism due to its widespread use when setting regulatory capital.[11] With numerous financial institutions using similar internal risk management models, it is possible that the constraint of VaR as the determinant of regulatory capital deepened the crisis. It is well known that during crisis periods, asset correlations can rise considerably, rendering portfolio diversification ineffective and causing numerous assets to decline in unison. As portfolio losses rise, the increase in VaR results in greater regulatory capital requirements, forcing financial institutions to sell asset positions in illiquid markets at fire sale prices. This has the potential to trigger a feedback effect where losses lead to further asset sales for the purpose of satisfying regulatory capital. This downward spiral can eventually lead to insolvency (Milne, 2008).

Another weakness is that different VaR methodologies can result in extremely different estimates for identical portfolios. Beder's (1995) study, which we introduced earlier, is a prime example. Beder applied eight different VaR methodologies and found that estimates could vary by more than 14 times for the same portfolio. Each approach has its own strengths and weaknesses, thereby making certain approaches superior depending on the situation. We earlier cited a study which found that 73% of banks use historical simulation and found that during periods of increased volatility, the approach produces downward biased estimates. One possible reason for the popularity of historical simulation among financial institutions is that it produces understated VaR estimates, especially during crisis periods. This in turn leads to banks having to post less regulatory capital. Banks may therefore have incentive to implement models which do not necessarily do the best job at describing firm-wide risks.

Some of the criticisms against the usefulness of Value at Risk in turbulent markets however, can be deemed to be unfair. VaR models use historical data in order to model potential future loss. It would be impossible (and not within the scope of the model) for a statistical measure which relies on historical data to accurately account for radical changes in volatility and asset correlations, or predict a major macroeconomic event such as a financial crisis. Therefore instead of over relying on VaR, prudent risk

[11] Nassim Taleb suggested to congress that they ban VaR due to its scientific uncertainty and because it gives traders a false sense of security.

management should supplement VaR with other risk management methodologies and more importantly, careful judgement.

10 Conclusion

The global financial crisis has led to increased scrutiny over risk management systems in financial institutions. This has called into question both the accuracy of Value at Risk estimates during periods of high volatility, and the merits of VaR as the determinant for regulatory capital. Although regulators have implemented the Stressed VaR to increase the Market Risk Charge when financial markets are unstable, the current regulatory framework is far from perfect due to the feedback effect along with other reasons highlighted in Section 8.

In this paper, we have analyzed four popular volatility models used to compute VaR in order to gauge whether their estimates can be relied upon in both stable and crisis periods. The theoretical section of this paper discussed these methodologies, identifying their underlying assumptions and shortcomings. Our results indicate that nonparametric approaches, particularly historical simulation, are inferior to parametric approaches as they tend to highly underestimate risks in crisis periods. This can be attributed to their inability to account for volatility clustering. Moreover, we conclude that models which assume t-distributed residuals outperform the normally distributed alternative as they are able to capture a greater degree of leptokurtosis and thus better model the observed return distribution. Although we find that the EGARCH model provides marginally stronger backtesting results compared to the ordinary GARCH model, we cannot conclude that one volatility model is clearly superior to the other. The key takeaway message to focus on is that the distributional assumption for returns plays a much larger role than the type of model being implemented. Therefore, of those models tested in this paper, we recommend GARCH or EGARCH models with t-distributed residuals as the best models for estimating VaR during the pre-crisis, and more importantly, crisis period**.** Lastly, we emphasize that Value at Risk is a useful risk management tool, primarily during normal market conditions, but should never be over relied upon. It is imperative that VaR be supplemented with other risk management methodologies and more importantly, careful judgement to mitigate future crises.

11 References

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1997). Coherent measures of risk. *Mathematical Finance*, 9(3), 203-228.

Bank of International Settlements (1996). Overview of the amendment to the capital accord to incorporate market risks.

Bank of International Settlements (2012). Fundamental review of the trading book.

Bartram, M. and Bodnar, M. (2009). No place to hide: The global crisis in equity markets in 2008/2009. *Journal of International Money and Finance*, 28(8), 1246-1292.

Becketti, S. (2013). *Introduction to Time Series Using Stata.* Stata Press books.

Beder, T.S. (1995). Var: Seductive but dangerous. *Financial Analysts Journal*, 12–24

Black, F. (1976). Studies of Stock Market Volatility Changes. *Proceedings of the American Statistical Association*, 177-181.

Bollerslev, Tim (1986) Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307–327.

Boudoukh, J., Richardson M., and R. Whitelaw (1998). The best of both worlds: A hybrid approach to calculating value at risk. *Risk*, 11(5), 64–67

Campbell, S. D. (2005). A review of backtesting and backtesting procedures. Divisions of Research & Statistics and Monetary Affairs. *Federal Reserve Board.*

Chen, J. (2013). Measuring Market Risk Under Basel II, 2.5, and III: VAR, Stressed VAR, and Expected Shortfall. Stressed VAR, and Expected Shortfall.

Christoffersen, P. (1998). Evaluating interval forecasts. *International economic review*, 841-862.

Christoffersen, P. (2003). *Elements of financial risk management*. Elsevier.

Davidson, R., & MacKinnon, J. G. (2004). *Econometric theory and methods*. New York: Oxford University Press.

Diebold, F. X., Schuermann, T., and Stroughair, J. D. (2000). Pitfalls and opportunities in the use of extreme value theory in risk management. *Journal of Risk Finance*, 1(2), 30-35.

Dowd, K. (1998). *Beyond Value at Risk: The New Science of Risk Management*. John Wiley and Sons Ltd.

Engle, J., and Gizycki M. (1999). Conservatism, accuracy and efficiency: Comparing value at risk models. *Working Paper: Australian Prudential Regulation Authority*

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of variance of UK inflation, *Econometrica*, 50, 987–1007.

Engle, R., and Manganelli, S. (2001). Value at risk models in finance. *European Central Bank Working Paper No.75.*

Engle, R., and Ng, K. (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5), 1749-1778.

Hull, J., and White A. (1998). Value at risk when daily changes in market variables are not normally distributed. *Journal of Derivatives*, 5(3), 9–19.

Jorion, P. (1997). *Value at risk: the new benchmark for controlling market risk.* New York: McGraw-Hill.

Kupiec, P.(1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2).

Lambert, P. and S. Laurent. (2001). Modeling Financial Time Series Using GARCH-Type Models and a Skewed Student Density. Mimeo, *Université de Liège.*

Milne, F. (2008). Anatomy of the credit crisis: The role of faulty risk management systems. *C.D. Howe Institure Commentary.*

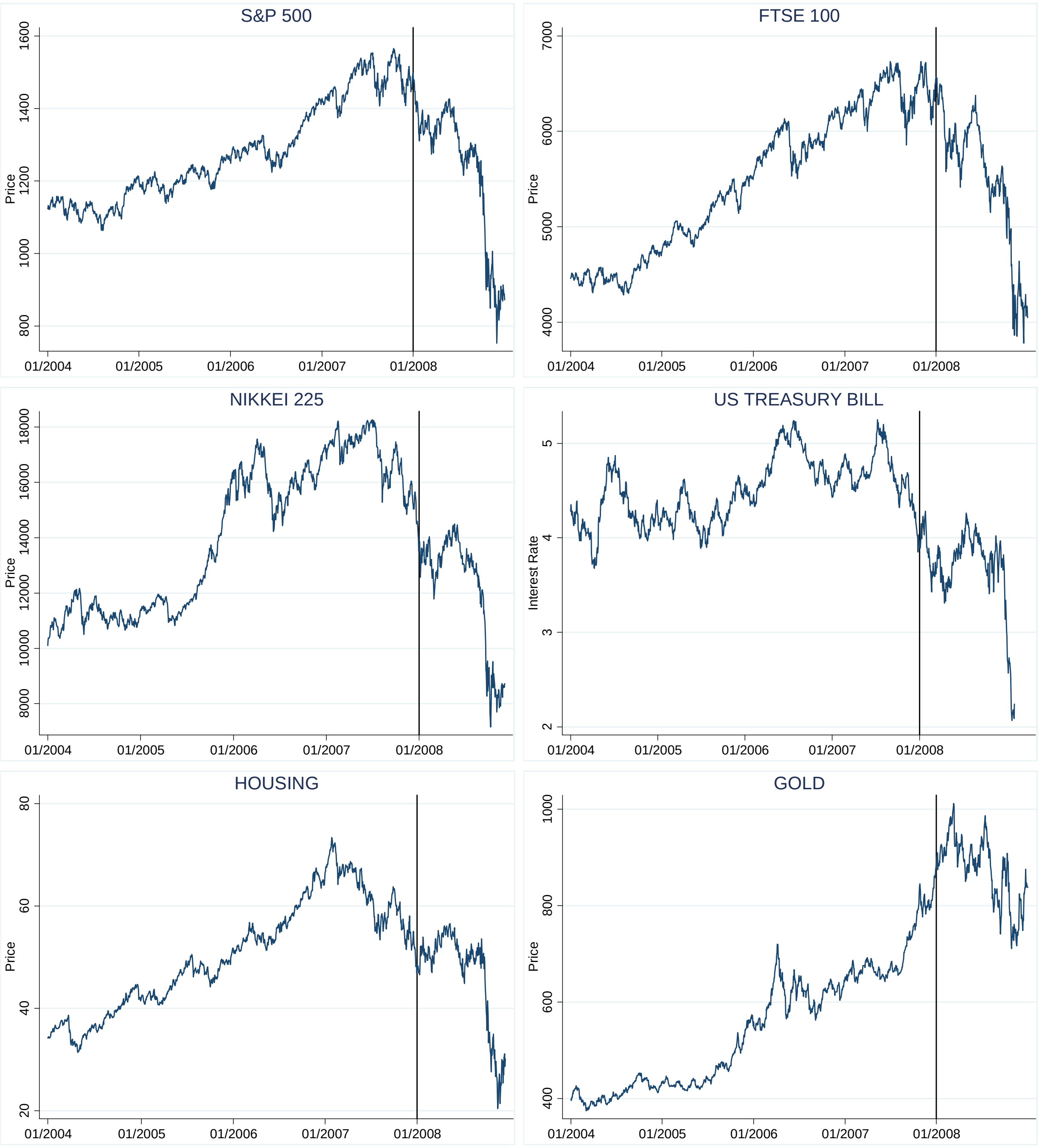
Nelson, B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, 59, 347–370.

Riskmetrics, T. M. (1994). JP Morgan Technical Document.

Wang, K. L., Fawson, C., Barrett, B., and McDonald, B. (2001). A flexible parametric GARCH model with an application to exchange rates. *Journal of Applied Econometrics*, 16(4), 521-536.

12 Appendix

Figure 1: Asset prices for pre-crisis and crisis periods.



Notes: The vertical black line represents the split between pre-crisis and crisis periods.

Figure 2: Histogram of empirical asset returns with an overlying normal distribution