ABSTRACT

NIHLANI, KIRAN. Nonparametric Estimation and Backtesting of Financial Risk Measures. (Under the direction of Peter Bloomfield.)

Measuring risk is an important question in finance, and the recent financial crisis has led to an overwhelming academic and regulatory response, so that we now have a theoretical corpus on risk measurement techniques. An important component in the risk measurement exercise is the evaluation of these measurements against the outcomes. This procedure is known as backtesting. Two measures of risk recommended by the Basel Committee on Banking Supervision (BIS) are Value at Risk (VaR) and Expected Shortfall (ES). VaR has been the preferred risk measure for its computational ease and simplicity. However, the lack of coherence and the inability to capture "tail risk" has led the regulators to shift to ES. The major challenge with the proposed change is the unavailability of efficient yet easy to implement tools to backtest ES.

The early VaR models were mostly parametric and imposed strong theoretical assumptions and rules on the underlying properties of the data. One such assumption is that the distribution of the data conforms to a known theoretical distribution. While these assumptions provided computational simplicity, the empirical evidence and VaR failures during the Mexico (1996), Asian (1997), and Russian (1998) market crises spurred the search for better VaR models. A number of non-parametric and semi-parametric VaR models were subsequently developed. The most commonly used of these include historical simulation, CAViaR, and extreme value analysis.There is a growing number of VaR measures based on quantile regression and its adaptations. Since VaR is a quantile of loss, it can be implicitly defined as a conditional quantile function making the quantile regression approach a natural environment to study and investigate VaRs. Merging the quantile regression approach with RiskMetrics, we get an exponentially weighted quantile regression framework that can be viewed in a kernel based framework. We proposed a modified way of implementing the exponentially weighted double kernel framework that improves the quality of the VaR forecasts and simplifies it for computational ease.

The quantile regression framework along with being a preferred base for VaR measures is a significant addition to the VaR and ES backtesting literature. A large portion of the VaR backtesting techniques including the recommended traffic-light based system by the Basel Committee on Banking Regulations are based on binary variables, such as whether or not there was an exception. This approach sacrifices too much information especially in finite samples. The quantile regression framework provides for a random coefficient model that can be used to perform a Wald type inference for the null hypothesis that the VaR model is correctly specified. Furthermore, since ES is the aggregated loss exceeding the VaR level, it can be approximated using VaR at a finite number of quantile levels. This helps propose a backtesting framework

for ES using a multi-quantile regression approach. We explore these backtests for VaR and ES through a comparative analysis with other backtests already established in the literature. The purpose here is to shed light on common issues underlying the implementation of these backtests and suggest modifications that provide significant improvements in computational ease and accuracy.

Nonparametric Estimation and Backtesting of Financial Risk Measures

by
Kiran Nihlani

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2020

APPROVED BY:

David Dickey	Sujit Ghosh
Denis Pelletier	Peter Bloomfield Chair of Advisory Committee

DEDICATION

To Ma, Papa, and Bhai.
For their endless love, support and encouragement.

BIOGRAPHY

The author was born on July 6, 1992 in Raipur, India. She completed her secondary education (till 10th grade) in 2008, and high school education in 2010. She went on to attend Banaras Hindu University where she was first introduced to Statistics, and graduated top of her class in the Bachelor of Arts degree examination in 2013. She continued her education at Savitribai Phule Pune University (previously University of Pune) in the Department of Statistics and received her Master of Arts in Statistics in 2015 before moving to North Carolina State University for her PhD in Statistics.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Peter Bloomfield, for his guidance, knowledge, and teaching. His support, dedication, and the ability to always provide me with positive, yet constructive feedback helped me immensely throughout the development of this dissertation. I am also grateful to him for his patience as our research was often slow because of my teaching commitments, and over-zealousness in completing professional certifications. I am also thankful for the insightful conversations that wandered from finance to world politics, and also Cricket. I am also thankful to Dr. David Dickey, Dr. Sujit Ghosh, and Dr. Denis Pelletier for serving on my committee and providing helpful feedback through this process.

I am also thankful to other NC State faculty and staff members who have assisted me throughout this process. First, thank you to Dr. Roger Woodard, Dr. Brooke Alhanti, and Dr. Dennis Boos for their mentoring during my teaching assignments. I owe special thanks to Dr. Vanessa Doriott Anderson, and Dr. Peter Harries for their encouragement and guidance during the completion of the Teaching and Communication Certificate, and Preparing the Professoriate program. I would also like to thank Dr. Cathrine Hoyo for giving me an opportunity to be a part of her Epidemiology and Environmental Epigenetics Lab. I have learnt a lot while working with her in different capacities, and I hope to continue our collaborations further. I am grateful to Dana Derosier for her friendship and faith that boosted my self-belief when things got tough. I also appreciate the support of Alison McCoy and Lanakila Alexander for not just their help regarding graduate school matters but also for all the restaurant and trip recommendations that made this journey more enjoyable.

My time in graduate school would not have been complete without the help and support of great friends in the department: Ajay Kumar, Indranil Sahoo, Suman Majumder, Jocelyn Chi, Meredith King, Vaidehi Dixit, and many others. I would also like to thank Gaurav Jain, Aditya Sawadh, and Anurag Singhal for their encouragement and for making sure my time outside the walls of SAS Hall was enjoyable. I am grateful for those enriching conversations over walks, coffee, and good food. I reserve special thanks for my closest friend, Nehal Jain, who has been a constant support and confidante for the last 12 years. Nehal Di, thank you for being your amazing self, and for taking me under you as your little sister.

Most importantly, I would like to thank my parents and my brother for their constant love and support. You have all worked hard to provide me with everything I needed to succeed. You instilled in me the values of hard work and sincerity that has served me well time and again throughout my years. You have always pushed me to challenge myself, and encouraged me to do what I love. Even on the toughest days, you encouraged me to keep going, reminded me to believe in my gifts, and to persevere to achieve my goals. Ma, you have been an amazing role model all my life. You taught me to dream big even when the means were sparse and

the journey difficult. Papa, you introduced me to the world through stories, newspapers, and encyclopedias. Bhai, your relentless perseverance along with the unabashed silliness inspires me to keep pushing forward without taking myself too seriously. I am grateful to have you all in my life.

Finally, I would like to appreciate all the essential workers who continue to provide essential services during this time of COVID-19. I have always had a huge respect for healthcare workers having seen my parents, and now my brother and sister-in-law dedicate their lives towards caring for others. I am in awe of their dedication, courage, and selflessness as we continue to work to overcome this challenge. I am grateful for your efforts, sacrifices, commitment, and leadership during these uncertain times. I aspire to learn, and take inspiration from them, and give back to the community to the best of my capabilities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

In the wake of recent financial crises, and the increasing complexity of the financial markets, an accurate assessment of risk exposure associated with an institution has become of paramount importance for both internal risk control, and financial regulation. A market risk measure summarizes the risk of losses due to movements in financial market variables into a single number. The variables include but are not limited to interest rates, foreign exchange rates, equity, and commodity prices. While financial firms take on a lot of market risk and thus reap the profits (and losses), they typically try to choose the type of risk they want to be exposed to. These risk measures are beneficial in many respects; however, they pose new questions like what risk measures are appropriate and how we can test their performance.

In the past, a variety of ad hoc tools were used to measure market risk like *notional amounts*, *sensitivity measures*, and *scenarios*. While these measures provided some intuition into the risk associated with a portfolio, they did not measure the downside risk. In addition, they failed to take into account the correlation across risk factors, and the probability of adverse movements of these factors. The notional amount only provided an indication of potential loss. Duration as a sensitivity measure was used to assess the impact of interest rates on bond prices but they failed to comment on whether a sharp movement in interest rates was likely. Scenario analysis provided some improvement over the other two methods by factoring in the nonlinear, extreme movements in price in the analysis. However, it did not associate probabilities with the losses. The notion of risk measurement has since evolved and become more sophisticated.

Risk measurement has two key components: a *risk metric* is the attribute of the risk being

measured, and *risk measure* is an operation that quantifies that risk. For instance, the *delta* of a stock option indicates the change in the option price for a unit change in the underlying stock price. Delta is the risk metric here, and it can be calculated using different risk measures. According to Holton (2004), risk entails two key components - uncertainty, and exposure, and risk metrics typically take one of the three forms where it either quantifies just the exposure or uncertainty, or both. The metrics that quantify uncertainty - by itself or in combination with exposure, are usually probabilistic.

Value at Risk (VaR) as a concept has been around since 1988 but it did not enter the financial lexicon until the early 1990s. It is now the most frequently used risk metric for market risk. It is defined as the maximum potential loss for a portfolio over a certain holding period at a given probability ("confidence") level.[1] For instance, if the estimated 99% 1-day VaR of a bank is 5 million then we can say with 99% confidence that bank will not lose more than 5 million dollars within the next day. A risk measure associated with quantifying the value at risk metric is called a value at risk measure. VaR metric is one of the probabilistic metrics of market risk available.

Even though VaR represents a simple concept, its measurement is a difficult statistical problem. There are many methods of calculating VaR depending on the input data, conditions, and complexity of calculations. These measures albeit using different methodologies have the same general structure: a) calculation of the present value of the portfolio, b) estimation of the distribution of changes in the portfolio, and c) the calculation of VaR. The key difference in the different VaR measures arise due to a difference in b).

There are different ways of classifying the VaR measures. First, they can be categorized into three groups on the basis of how the VaR is calculated - using historical simulation, using an analytical approach, or by using a Monte Carlo simulation. These measures can be alternately classified as a) Delta method that includes different variants of the analytical methods, and b) Monte Carlo methods that encompass all simulation based methods including historical simulation. Another way of classifying these measures is based on the underlying distributional assumptions used to calculate the VaR as: a) parametric, b) nonparametric, and c) semiparametric. This classification based on distributional assumptions is discussed in more detail in Chapter 2.

The regulatory framework put forward by Basel Committee III requires the banks to report their 99% VaR on a daily basis. The VaR reported specifies the amount of capital required to maintain this amount of risk for the bank, also called the capital charge. To ensure the validity of these VaR estimates, the numbers are backtested against the realized losses by counting the number of exceedances, i.e. the number of days a bank's loss exceed the VaR in a given period.

[1] In finance, the convention is to refer to the probability as a *confidence level*; this is uncomfortable for a statistician, but we shall follow the finance convention.

The supervisory committee framework entails formal testing and accounting of exceptions on a quarterly basis using the most recent twelve months of data. The committee uses a three-zone "traffic light" approach to assess the quality or accuracy of a bank's internal model.

The popularity of VaR is because of the ease to understand and implement it. Furthermore, it is simple to validate or backtest VaR by comparing the predicted risk with the actual risk. For instance, there is something wrong with our 99% VaR calculations if we observe 10 losses over 5 million for the bank in past 100 days. The biggest criticism of VaR has been its inability to capture tail risk. While it does provide a threshold for loss on a bad day, it gives no insight into the amount by which the loss would exceed the threshold. For instance, even if two distributions have the same VaR, they might not be equally 'risky' as shown in the figure below.

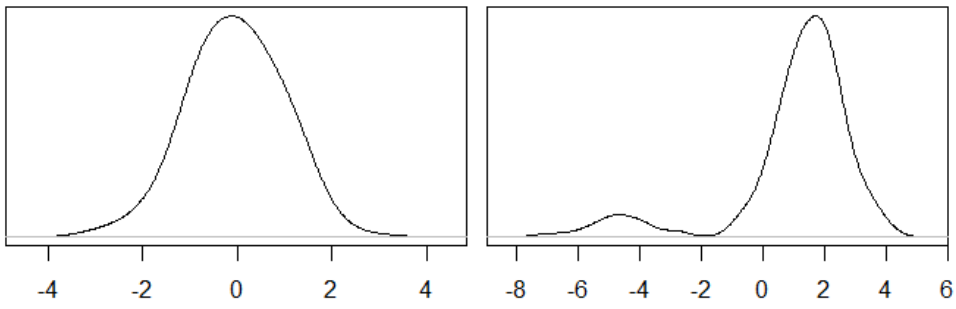


Figure 1.1 Two return distributions with the same 99% VaR = 2.33. Even when the VaR is same, the plot on the right depicts a more risky portfolio

The other criticisms of VaR include its lack of subadditivity. This means that the VaR for two or more portfolios combined can be greater than the sum of VaR for each of these portfolios independently. It suggests that diversification increases risk which is a contradiction of standard financial beliefs. The inability of VaR to capture tail risk and lack of subadditivity has led the supervisors to shift to another popular risk metric known as the *expected shortfall (ES)*. It is defined as the expected loss in the tail distribution of returns. In other words, it is the conditional average of losses that exceed a chosen quantile. Further, ES is also a coherent risk metric (Artzner et al. (1999)), which means it meets the subadditivity criterion thereby making it a preferred risk metric over VaR. In Figure 1.1, while the VaR is same for both figures, the ES is higher for the right-hand plot capturing the riskiness of the portfolio.

Traditionally, VaR has been used by the banking industry, and ES by the insurance industry. This is rapidly changing, and the criticisms of VaR have led the Basel Committee to propose that the 97.5% expected shortfall to be reported along with the 99% VaR. This gives a better account for the tail risk. While expected shortfall has many advantages over VaR, there is

one drawback that has prevented the supervisors from using ES as the sole risk measure, and that is *backtesting*. For any given risk measure, it is important that we are able to estimate it accurately, and also validate these estimates by checking whether the *ex ante* estimates or forecasts are in line with the realized losses observed *ex post*. The statistical procedure used to perform these comparisons is known as backtesting.

The literature on backtesting VaR estimates is large. It is based on the idea that if the VaR model is correctly specified at confidence level τ, the VaR exceptions, i.e. when realized losses exceed the VaR forecasts, form a sequence of independent, identically distributed (i.i.d) Bernoulli variables with probability $1 - \tau$. Kupiec (1995) proposed a test of *unconditional coverage* which is basically a binomial test for the number of VaR violations. A *conditional coverage* test that along with the binomial test also explicitly examines the independence of these violations was proposed in Christoffersen (1998) and Christoffersen et al. (2001). Further, a regression based coverage test was developed by Engle and Manganelli (2004) for checking the fit of their CaViaR model for dynamic quantiles.

The literature on backtesting ES is much smaller. There are several questions surrounding backtesting expected shortfall ranging from how to do it to whether it is even possible. Gneiting (2011) published in his work that expected shortfall lacks a property called "elicitability" that VaR has and is important for backtesting a risk measure. Recent studies however have shown that elicitability is required for model selections or comparing different methods with respect to expected shortfall but not to backtest them independently. McNeil and Frey (2000) suggested a bootstrap test based on violation residuals that measure the discrepancy between the ES forecast and the realized losses when a violation occurs. Acerbi and Szekely (2014) suggest three different statistics also based on such discrepancies using the Monte Carlo hypothesis tests. Du and Escanciano (2017) and Costanzino and Curran (2015) have recently proposed using a Z-test based on cumulative violations over several probability levels that can be used to approximate the ES. While transitioning from VaR to ES, the Basel Committee III maintained that while banks are expected to report ES at 97.5% but it shall be backtested based on the 99% VaR. Costanzino and Curran (2018) recently put forward a traffic-light system for ES similar to the traffic-light system used by the Basel Committee for backtesting VaR.

1.1 Outline and contribution

In Chapter 2, we build on the exponentially weighted quantile regression (EWQR) model presented by Taylor (2007) and propose a modified approach for implementing their exponentially weighted double kernel quantile regression technique (EWDKQR) with leverage effect. Taylor (2007) view the EWQR framework as a double kernel quantile regression by incorporating kernel density estimation within the EWQR framework. One of the estimation methods considered in

their text to use a linear quantile model for VaR with regressors to account for the leverage effect. Including regressors in the linear quantile model requires the quantile regression parameters to be dynamically updated at each iteration of the optimization problem for each time point. We propose an alternative three parameter optimization problem that now optimizes the slope parameter globally. We empirically show that doing so is not only computationally parsimonious but also yields better VaR forecasts as confirmed by different backtesting techniques.

In Chapter 3, we explore a recently proposed quantile regression based backesting technique for VaR. The backtest includes a quantile regression based hypothesis test that admits a Wald-type test. The asymptotics underlying these tests are sufficient for more central quantile levels, and therefore need modifications implement them effectively at extreme quantiles as in the case of 95% and 99% VaR. We provide a working guideline on using the sparsity based Hendricks-Koenker sandwich matrix with a scaled up bandwidth to allow for more observations, and therefore more information for calculating the test statistic for improved results. We present a Monte Carlo simulation study to compare the performance of the quantile regression based backtest with Hit and DQ tests (discussed in Chapter 2 in two different market scenarios).

In Chapter 4, we explore four different approaches for backtesting expected shortfall. We follow the idea proposed by Emmer et al. (2015) that ES can be approximated by VaR at several risk levels. This reduces the ES backtesting problem to backtesting VaRs at these risk levels. We explore a multi-quantile extension of the quantile regression based backtest used for VaR in Chapter 3 to jointly backtest VaR at these levels. We also present a simulation study to compare two approaches for implementing these tests by White et al. (2015) and Koenker (2005) across two different market scenarios.

Based on Chapters 2-4, Chapter 5 briefly discusses potential areas of future work. Chapters 2-3 discuss VaR, its estimation, and backtesting. We conclude the introduction with a brief overview of mathematical properties of risk measures, and two backtesting techniques that are well established in the literature, and commonly used for backtesting VaR.

1.2 Mathematical properties of risk measures

There are many ways in which the risk can be summarized by a single number. The most commonly used of these is the standard deviation. A commonly used measure of systemic risk is Beta (β or beta coefficient). However, while risk is mainly concerned with losses, both standard deviation and beta coefficient as measures of volatility take measure deviations both up and down. There are certain properties that a good risk measure must have. Artzner et al. (1999) proposed a set of criteria that a measure of risk $\rho(X)$, where X is a portfolio, should satisfy. These criteria are crucial in understanding the academic debate on the differences between VaR

and ES.

A risk measure is said to be *coherent* if it has the following properties:

- *Subadditivity:* The risk measure of two portfolios after they have been merged cannot be any worse than adding the two risks separately. Mathematically, for two portfolios X and Y,

  $$\rho(X + Y) \leq \rho(X) + \rho(Y)$$

  Subadditivity is an obvious requirement for a risk measure, without which there would be no benefit from adding uncorrelated new trades. Lack of this property can lead to a form of regulatory arbitrage- a bank can create many subsidiary firms to save regulatory capital.

- *Monotonicity:* If a portfolio X produces a worse result than another portfolio Y for every scenario, its risk measure should be greater, i.e.,

  $$\text{If } X < Y, \quad \text{then} \quad \rho(X) > \rho(Y)$$

- *Positive homogeneity:* Changing the portfolio by a factor of λ while keeping the relative amounts of different items in the portfolio the same, should result in the risk measure being multiplied by λ. For a portfolio X,

  $$\rho(\lambda X) = \lambda \rho(X) \ , \quad \lambda > 0$$

- *Translation invariance:* If an amount of cash is added to the portfolio, its risk measure should go down by the amount. If cash amount c is added to the portfolio X,

  $$\rho(X + c) = \rho(X) - c$$

Artzner et al. (1999) pointed out that VaR is not a coherent risk measure because of the lack of subadditivity. Expected shortfall on the other hand satisfies all four properties and is therefore a coherent risk measure.

1.3 Value at risk

We now present a formal definition of VaR. The VaR of a portfolio at confidence level $\tau \in (0, 1)$ is the smallest number x such that the probability that the loss X exceeds x is no larger than

$1-\tau$. Mathematically,

$$\text{VaR}_\tau(X) = \inf\{x \in \mathbb{R} : P(X > x) \leq 1-\tau\} \tag{1.1}$$

$$= \inf\{x \in \mathbb{R} : F_X(x) \geq \tau\} \tag{1.2}$$

where $F_X(.)$ is the loss distribution associated with the portfolio. The above expression implies that $\text{VaR}_\tau(X)$ is just the τ-th quantile of X. It should be noted that VaR is a function of two parameters- the confidence level τ, and the time horizon or the holding period for the portfolio, T. Expression (1.1) assumes the time horizon, $T = 1$ day. VaR satisfies three of the four properties of a coherent risk measure viz. monotonicity, positive homogeneity, and translation invariance. A simple example is present here to show that VaR is not subadditive.

Example 1.3.1. Consider two independent projects, each with a 0.02 probability of losing \$5 million and a 0.98 probability of losing \$1 million. The 97.5% sum of the VaR for the projects individually is \$ 2 million. Suppose a new portfolio is constructed by combining both projects. The portfolio has a $0.02 \times 0.02 = 0.0004$ probability of losing \$10 million, a $2 \times 0.02 \times 0.98 = 0.0392$ probability of losing \$6 million and a $0.98 \times 0.98 = 0.9604$ probability of losing \$2 million. A 97.5% VaR of the combined portfolio is \$6 million which is \$4 million more than the sum of the VaRs for the projects independently. This shows that VaR is not subadditive. It should be noted that while these projects would not be found at any institution since they only lose money, they can be modified to be profitable without changing their VaRs.

Definition 1.3.1. Expected shortfall is the average of losses exceeding the value at risk. Mathematically, it can be written as:

$$\text{ES}_\tau(X) = \frac{1}{1-\tau} \int_\tau^1 \text{VaR}_u(X)\, du \tag{1.3}$$

Expected shortfall is also known by other names including conditional value at risk (CVaR), average value at risk (AVaR), and expected tail loss (ETL).

1.4 Elicitability

The concept of elicitability was introduced by Osband and Reichelstein (1985), further developed by Lambert et al. (2008), and applied in the context of risk measures by Gneiting (2011). The two key elements of financial risk measurement are the choice of a suitable risk measure, and of a forecasting method. Guidelines on choosing a risk measure have been discussed in a pro-con framework by Emmer et al. (2015), as an approach to consider the entire "risk-measurement procedure" (rather than just the choice of the risk measure) by Cont et al. (2010), and consistency of risk measures by Davis (2016).

The performance of a risk measure over time can be evaluated by comparing the realized losses with the risk measure forecasts. This process is known as backtesting. Traditional backtesting measures are constructed to test the null hypothesis that the risk measure is adequate. The three-zone "traffic light" approach by the Bank of International Settlements (BIS) is one such backtesting technique. Other commonly used backtesting methods for VaR are by Christoffersen and Pelletier (2003), McNeil and Embrechts (2005) and Engle and Manganelli (2004). While these traditional backtesting methods assess the optimality of risk measures, they do not provide a comparative evaluation between them. The property of elicitability provides insights on performing these comparative evaluations by using scoring functions.

1.4.1 Definition

Let $x_1, \ldots, x_n$ be the realized observation values and $y_1, \ldots, y_n$ be the corresponding forecasts. A performance criterion is then given as:

$$\bar{S} \;=\; \frac{1}{n}\sum_{i=1}^{n} S(x_i, y_i)$$

The function S, called the *scoring function*, depends on both the forecasts and the realized observations. Table 1.1 lists some commonly used scoring functions. A detailed analysis of these and many other scoring functions can be found in Patton (2009).

Table 1.1 Some commonly used scoring functions

$S(x,y) = (x-y)^2$	squared error (SE)
$S(x,y) = \lvert(x-y)\rvert$	absolute error (AE)
$S(x,y) = \lvert(x-y)/y\rvert$	relative error (RE)
$S(x,y) = \lvert(x-y)/x\rvert$	absolute percentage error (APE)

A forecasting statistic that can be expressed as a minimized value of a scoring function is said to have the property of elicitability. A statistic ψ is said to be elicitable if it is the minimized value of some scoring function $S(x, y)$:

$$\psi \;=\; \arg\min_{y} \mathbb{E}_F[S(X, y)]$$

where F is the distribution of the realized observations. This distribution can be empirical, parametric, or simulated. Let $\mathbb{I}$ be the potential range of outcomes, and let F be the probability distribution function concentrated on $\mathbb{I}$. Then a scoring function is any mapping $S : \mathbb{I}\times\mathbb{I} \to [0, \infty)$. A *functional* is a set-valued mapping defined as $F \to T(F) \;\subseteq\; \mathbb{I}$. A scoring function S is

consistent for the functional T if

$$\mathbb{E}_F[S(X,t)] \;\leq\; \mathbb{E}_F[S(X,y)]$$

for all F, all $t \in \mathrm{T}(F)$, and all $y \in \mathbb{I}$. It is *strictly consistent* if it is consistent, and the equality of the expectations implies that $x \in \mathrm{T}(F)$. Further, a functional is *elicitable* if there exists a strictly consistent scoring function for it. Some examples for a univariate predictand include expectations, quantiles, ratios of expectations and expectiles.

1.4.2 Elicitability of VaR

Gneiting (2011) show that for a strictly increasing function g, the scoring function

$$S(x,y) \;=\; (I_{\{y \geq x\}} \;-\; \tau)\,(g(y) \;-\; g(x)) \tag{1.4}$$

is consistent for the τ-quantile. However, it is not the most widely used scoring function for VaR forecasts, or quantile risk measure in general. The most widely recommended scoring function is the piece-wise linear scoring function which basically is the same scoring function as in (1.4) with $g : \mathbb{R} \to \mathbb{R}$ as an identity function. The piece-wise linear scoring function, $S_{PL} : \mathbb{R} \to [0, \infty)$ is given by:

$$S_{PL}(x,y) \;=\; (I_{\{y \geq x\}} \;-\; \tau)\,(y \;-\; x) \tag{1.5}$$

To prove the elicitability of VaR, we need to show that:

$$\mathrm{VaR}_\tau(X) \;=\; \arg\min_y \mathbb{E}[(I_{y \geq x} \;-\; \tau)\,(y \;-\; x)] \;= F_X^{-1}(\tau) \tag{1.6}$$

Using the Heavside step function, $\theta(y-x)$, that takes the value one for positive arguments and zero for negative arguments as $\theta(y-x) \;=\; I_{y \geq x}$, the scoring function can be written as:

$$S(x,y) \;=\; (\theta(y-x) \;-\; \tau)\,(y \;-\; x)$$

The expected value of the scoring function can then be written as

$$\begin{aligned}
\mathbb{E}[S(x,y)] \;&=\; \mathbb{E}\left[(\theta(y-x) \;-\; \tau)\,(y \;-\; x)\right] \\
&=\; \int_{\infty}^{\infty} (\theta(y-x) \;-\; \tau)\,(y \;-\; x)\, f_X(x)\, dx \\
&=\; (1-\tau) \int_{-\infty}^{y} (y \;-\; x)\, f_X(x)\, dx \;-\; \tau \int_{y}^{\infty} (y \;-\; x)\, f_X(x)\, dx
\end{aligned}$$

We now equate the first derivative of $\mathbb{E}[S(x,y)]$ to 0 and solve for y. Taking the derivative of

the first term in the expression above:

$$\frac{d}{dy}\left((1-\tau)\int_{-\infty}^{y}(y\ -\ x)\ f_X(x)\ dx\right)\ =\ (1-\tau)\int_{-\infty}^{y}f_X(x)\ dx$$

Similarly, taking derivative of the second term,

$$\frac{d}{dy}\left(-\tau\ \int_{y}^{\infty}(y\ -\ x)\ f_X(x)\ dx\right)\ =\ -\tau\ \int_{y}^{\infty}f_X(x)\ dx$$

Adding the two derivative terms, we get

$$\begin{aligned}\frac{d}{dy}\mathbb{E}[S(x,y)]\ &=\ (1-\tau)\int_{-\infty}^{y}f_X(x)\ dx\ +\ -\tau\ \int_{y}^{\infty}f_X(x)\ dx\\ &=\int_{-\infty}^{y}f_X(x)\ dx\ -\ \tau\end{aligned}$$

Setting the above expression equal to zero we have,

$$\tau\ =\ \int_{-\infty}^{y}f_X(x)\ dx\ \implies\ y\ =\ F_X^{-1}(\tau)$$

The second derivative is positive implying that VaR in fact minimizes the scoring function and is elicitable.

1.4.3 Elicitability and backtesting

Gneiting (2011) showed that expected shortfall is not elicitable. This means there does not exist a scoring function such that ES minimises it. This result shifted the debate on how to backtest expected shortfall to whether it was even possible to do so. Following up on this argument, Emmer et al. (2015) showed that ES is conditionally elicitable, consisting of two elicitable components. This meant that ES could be backtested by testing the two components separately. They suggest the following two-step procedure to forecast ES:

- Calculate the quantile as

$$\mathrm{VaR}_\tau(X)\ =\ \arg\min_y\ \mathbb{E}\left[(I_{y\geq x}\ -\ \tau)\ (y\ -\ x)\right]$$

- Calculate $\mathrm{ES}_\tau(X)\ =\ \mathbb{E}\ [L|L\ \geq\ \mathrm{VaR}_\tau]$, where $L = -X$ is the loss using the scoring function $\mathbb{E}_P[(y-X)^2]$, with probabilities $\tilde{P}(A) = P(A|L \geq \mathrm{VaR}_\tau)$. This gives

$$\mathrm{ES}_\tau(X)\ =\ \arg\min_y\ \mathbb{E}_{\tilde{P}}\ [(y\ -\ X)^2]$$

Since VaR is elicitable, the only thing remaining is the conditional expectation and expectations are elicitable.

Acerbi and Szekely (2014) argue that ES is backtestable even without using the conditional elicitability condition. They reiterate that elicitability of a measure is useful for relative comparison of different models but not relevant for an absolute model evaluation. Even though VaR is elicitable, there are very few methods for estimating VaR that employ this property. They propose three non-parametric ES backtest methods that are easy to implement and have higher power than the standard Basel VaR backtests.

1.5 Backtesting VaR

Backtesting, in simple words, is testing a predictive model based on the historical data. Jorion (2007) defines backtesting as a set of statistical procedures designed to check if realized losses were in line with the VaR forecasts. This definition can further be extended to any risk measure. The choice of the backtesting method depends on the type of forecast- point forecasts, interval forecast, or forecasts for the complete probability distribution. Backtesting methods for VaR started appearing in the literature as VaR gained prominence as a risk measure. The earliest backtesting techniques can be found in Kupiec (1995) and Hendricks (1996).

1.5.1 Coverage tests

Christoffersen (2008) devised the popular backtesting method for VaR based on the *violation process* or the *hit sequence*. Consider a continuous loss distribution. By definition of VaR at confidence level τ:

$$P(X > \text{VaR}_\tau(X)) = 1 - \tau$$

The violation process can then be defined as

$$I_t(\tau) \;=\; \mathbb{1}_{\{X_t > VaR_\tau(X_t)\}} \tag{1.7}$$

The hit sequence takes the value 1 if the loss on day t is larger than the value of VaR, and 0 otherwise. To backtest a risk model, we construct a sequence of hits $\{I_t\}_{t=1}^{T}$ indicating past violations across T days. Assuming the hit sequence of violations is completely unpredictable and therefore distributed independently over time as a Bernoulli variable, we have the VaR backtesting setup as:

$$H_0 :\; I_t \sim \text{ i.i.d. Bernoulli}(1-\tau)$$

More precisely, this can be broken down to two different hypothesis corresponding to conditional and unconditional coverage.

1.5.1.1 Unconditional Coverage Test

A risk model for VaR estimates has the correct unconditional coverage if

$$P(I_t = 1) = \mathbb{E}\left[I_t\right] \;=\; 1 - \tau$$

and the unconditional coverage hypothesis can be written as:

$$H_0 :\; \mathbb{E}\left[I_t\right] = 1 - \tau$$

The unconditional coverage test evaluates if the unconditional probability of violation under the assumed risk model is different from the expected probability $p = 1 - \tau$. It is sufficient to verify if the number of violations follow a binomial distribution:

$$f(x) \;=\; \binom{T}{x} p^x (1-p)^{T-x}$$

Normal approximation can also be used in case of a large number of observations.

1.5.1.2 Conditional Coverage Test

A risk model has the correct conditional coverage if

$$P_t(I_{t+1} = 1) = \mathbb{E}_t\left[I_{t+1}\right] \;=\; 1 - \tau$$

where P_t is the conditional distribution given information until time t. The conditional coverage hypothesis can be written as:

$$H_0 :\; \mathbb{E}_t\left[I_{t+1}\right] = 1 - \tau$$

The unconditional coverage tests are not useful in situations where violations are clustered in time. For instance, if a 5% VaR gave exactly 5% violations but all clustered within a 2-week period, the risk is much higher than if these violations were scattered randomly. Conditional coverage tests jointly test that the VaR violations are independent and the average number of violations is correct.

Assume that the hit sequence is dependent over time and that it can be described as a

first-order Markov sequence with transition probability matrix

$$\Pi_1 \;=\; \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}$$

where π_{01} denotes the probability of a violation tomorrow conditional on today being a non-violation and π_{11} is the probability of tomorrow being a violation given that today is also a violation. If we observe a sample of T observations, the likelihood function of the first-order Markov process can be written as

$$L(\Pi_1) \;=\; (1-\pi_{01})^{N_{00}}\, \pi_{01}^{N_{01}}\, (1-\pi_{11})^{N_{10}}\, \pi_{11}^{N_{11}}$$

where $N_{ij},\ i,\ j \;=\; 0,\ 1$ is the number of observations with a j following an i. Let $\hat{\Pi}_1$ be the matrix of estimated transition probabilities, i.e.,

$$\hat{\Pi}_1 \;=\; \begin{bmatrix} 1-\hat{\pi}_{01} & \hat{\pi}_{01} \\ 1-\hat{\pi}_{11} & \hat{\pi}_{11} \end{bmatrix}$$

where $\hat{\pi}_{01}$ and $\hat{\pi}_{11}$ are maximum likelihood estimates. The conditional coverage test is then given as

$$\mathrm{LR}_{\mathrm{cc}} \;=\; -2\ln\left[L(p)\;/\;L(\hat{\Pi}_1)\right] \;\sim\; \chi^2_2$$

which corresponds to testing that $\pi_{01} \;=\; \pi_{11} \;=\; p$. The likelihood ratio test statistic $\mathrm{LR}_{\mathrm{cc}}$ can also be written as:

$$\begin{aligned}
\mathrm{LR}_{\mathrm{cc}} \;&=\; -2\ln\left[L(p)\;/\;L(\hat{\Pi}_1)\right] \\
&=\; -2\ln\left[\{L(p)\;/L(\hat{\pi})\}\;\left\{L(\hat{\pi})\;/\;L(\hat{\Pi}_1)\right\}\right] \\
&=\; -2\ln\left[L(p)\;/L(\hat{\pi})\right] - 2\ln\left[L(\hat{\pi})\;/\;L(\hat{\Pi}_1)\right] \\
&=\; \mathrm{LR}_{\mathrm{uc}} \;+\; \mathrm{LR}_{\mathrm{ind}}
\end{aligned}$$

where $\mathrm{LR}_{\mathrm{uc}}$ is the likelihood from the unconditional coverage test and $\mathrm{LR}_{\mathrm{ind}}$ is the likelihood for the independence hypothesis given by

$$\mathrm{LR}_{\mathrm{ind}} \;=\; -2\ln\left[L(\hat{\pi})\;/\;L(\hat{\Pi}_1)\right] \;\sim\; \chi^2_1$$

1.5.2 Backtesting with information variables

While the tests mentioned so far are simple and easy to implement, Monte Carlo studies have shown that they do not have much power to detect an incorrect risk model. This is primarily because these models are based on only the past VaR violations. Christoffersen (2008) proposed

to take into account the information from past relevant market variables. This helps widen the information set, and better explain when and why a violation occurs.

Let Y_t denote the q-dimensional vector of market variables available at time t. The null hypothesis of this new model can be written as

$$H_0 \ : \ P(I_t \ = \ 1|\ Y_{t-1}) \ = \ p \ \Leftrightarrow \ \mathbb{E}\left[I_t - p|Y_{t-1}\right] \ = \ 0$$

This implies that the conditional probability of having a VaR violation on day t should be independent of the market variables observed at time t and coincide with the expected VaR coverage rate, $p = 1 - \tau$. The conditional expectation of the hit sequence is therefore equal to p.

Engle and Manganelli (2004) adopted a quantile regression framework and proposed the dynamic- Quantile (DQ) test for the null hypothesis:

$$DQ \ = \ \frac{(I-p)'Y(Y'Y)^{-1}Y'(I-p)}{Tp(1-p)} \ \sim \ \chi^2_q$$

where Y is a $T \times q$ matrix of market variables. Berkowitz and O'Brien (2002) have further extended the DQ framework by considering lagged VaR from a GARCH model and lagged violations. They point out that doing so gives a higher power for the test and also better VaR estimates.

Given this background information, we proceed to Chapters 2-4.

CHAPTER

2

VAR ESTIMATION USING EXPONENTIALLY WEIGHTED QUANTILE REGRESSION

2.1 Introduction

We know that estimation of VaR amounts to forecasting tail quantiles conditional on the information in hand. Several approaches have been proposed to estimate tail quantiles but there is no single established method. Engle and Manganelli (2004) divide these methods into three categories- parametric, semiparametric, and nonparametric. *Parametric* methods involve parametrization for the behavior of prices. Conditional quality forecasts are then constructed following a conditional volatility forecast and a distributional assumption. Exponential smoothing (RiskMetrics) or GARCH model is used to forecast the volatility and a Gaussian/Student-t distribution is assumed. While the parametric methods do have the advantage of allowing a complete characterization of distribution of returns, the normality assumption of the residuals doesn't seem consistent with the behavior of financial returns.

Semiparametric approaches to VaR calculations include methods based on extreme value theory like McNeil and Frey (2000) which involves applying the peak over threshold EVT method to standardized residuals from a GARCH conditional volatility estimates. This approach also

includes methods based on quantile regression like the conditional autoregressive value at risk (CAViaR) of Engle and Manganelli (2004). Other methods under this category include the quasi-maximum likelihood GARCH and EVT based CAViaR.

The most common *nonparametric* technique for VaR estimation is the historical simulation method which estimates VaR as the quantile of the empirical distribution of returns in a moving window of the most recent periods. The method involves an implicit assumption that the distribution of returns doesn't change within the window, which is logically inconsistent with the idea that the time series returns are not independent and identically distributed (i.i.d.). Further, there is no clear specification on the length of the moving window- it should be large enough to draw significant statistical inference but not so large that it includes observations outside the current volatility cluster. An alternative technique to overcome this issue was suggested by P. Richardson et al. (1997) where they propose using exponentially decreasing weights for past returns in a quantile regression framework.

Adding to the literature on nonparametric methods, Taylor (2007) proposes the use of exponentially weighted quantile regression (EWQR) to estimate VaR and the corresponding ES. The EWQR method basically is the same as the exponential smoothing of the cumulative distribution function (cdf) and can be viewed in either a single or a double kernel framework. VaR estimation in this setup can be done using EWQR with only an intercept and no regressors, but a regressor can be included to capture the leverage effect, if any.

2.2 Exponential weighted quantile regression

Quantile regression for linear models was introduced by Koenker and Bassett (1978),

$$Q_t(\tau) = y_t' \beta \tag{2.1}$$

where y_t is a vector of regressors, and β is a parameter vector. The quantile regression minimization problem is then given as:

$$\min_{\beta} \sum_{t=1}^{T} (x_t - y_t' \beta)(\tau - \mathbb{1}(x_t < y_t' \beta)) \tag{2.2}$$

where x_t is the centered log returns. The centering is done with respect to the conditional mean, which is often assumed to be zero or constant. EWQR extends this framework to incorporate an exponential weighting factor for the returns. For a given weighting parameter λ, the EWQR

minimization can be written as:

$$\min_{\beta} \sum_{t=1}^{T} \lambda^{T-t}(x_t - y_t' \beta)\,(\tau - \mathbb{1}(x_t < y_t' \beta)) \tag{2.3}$$

For a linear quantile model, this minimization in equations (2.2) and (2.3) can be formulated and solved as a linear program. Following Koenker (2005) the expected shortfall for a zero mean residual term x_t can be written as:

$$\text{ES} = \frac{1}{\tau} E((x - Q(\tau))(\tau - \mathbb{1}(x < Q(\tau)))) \tag{2.4}$$

This expression involves the expectation of the asymmetric "tick" function used in quantile regression in expression (2.2). The tick function in expression (2.4) is evaluated at the quantile $Q(\tau)$ which is estimated using the quantile regression minimization. A sample estimator of the expectation can then be written as:

$$\widehat{\text{ES}} = \frac{1}{\tau T} \sum_{t=1}^{T} (x_t - y_t' \hat{\beta})(\tau - \mathbb{1}(x_t < y_t' \hat{\beta})) \tag{2.5}$$

Quantile regression therefore gives us both an estimator for the quantile and an ES estimator. The latter however is of limited use because in view of the time-varying nature of financial returns we need an estimator for ES that is time-varying too. In other words, expression (2.5) provides an unconditional ES estimate while we need a conditional estimator based on information up to the current period. The EWQR setup can help overcome this problem by giving as an exponentially weighted time-varying ES estimator. Replacing the quantile regression objective function with the EWQR objective function in expression (2.4), the ES estimator can be written as:

$$\widehat{\text{ES}}_T = \frac{1}{\tau \sum_{t=1}^{T} \lambda^{T-t}} \sum_{t=1}^{T} \lambda^{T-t}\,(x_t - y_t' \hat{\beta})\,(\tau - \mathbb{1}(x_t < y_t' \hat{\beta})) \tag{2.6}$$

For a τ quantile in the upper tail of the distribution, the analogous expression is:

$$\widehat{\text{ES}}_T = \frac{1}{(1-\tau) \sum_{t=1}^{T} \lambda^{T-t}} \sum_{t=1}^{T} \lambda^{T-t}\,(x_t - y_t' \hat{\beta})\,(\tau - \mathbb{1}(x_t < y_t' \hat{\beta})) \tag{2.7}$$

2.2.1 Viewing EWQR as a kernel quantile regression

Jones and Hall (1990) consider a kernel weighting scheme for nonparametric estimation of

quantiles of x_t conditional on a scalar y_t:

$$\min_q \sum_{t=1}^{T} K_h(y - y_t)(x_t - q)(\tau - \mathbb{1}(x_t < q)) \tag{2.8}$$

where $K_h(y - y_t)$ is a kernel weighting function. The above expression uses only the intercept q with no regressors, implying a local linear fitting. It can be further shown that the standard kernel estimator for a CDF at a specified value x in period T is given as:

$$\widehat{F}_T(r) = \frac{1}{\sum_{t=1}^{T} K_h(y - y_t)} \sum_{t=1}^{T} K_h(y - y_t)\, \mathbb{1}(x_t < x) \tag{2.9}$$

Let $y_t = t$ and the location $y = T$ and define the kernel function to be a one-sided exponentially declining weight on data left of the location T, $K_h(y - y_t) = \lambda^{T-t}$. Substituting this in equation (2.8) gives us the EWQR minimization setup (only intercept, no regressors). This shows that EWQR can be viewed as a form of kernel quantile estimation.

One of the crucial issues when using kernel estimators is the choice of bandwidth. For the exponentially weighted kernel setup described above, this implies the choice of λ. A relatively smaller value of λ is needed for swift adaptation, if the distribution of returns is changing quickly. However, a larger value of λ is needed for estimating tail quantiles in order to give sizeable weights to observations. Based on the double kernel CDF estimator of Yu and Jones (1998), kernel density estimation can be incorporated within the above EWQR CDF framework:

$$\widehat{F}_T(x) = \frac{1}{\sum_{t=1}^{T} K_{h_1}(y - y_t)} \sum_{t=1}^{T} K_{h_1}(y - y_t)\, \Omega_{h_2}(x - x_t) \tag{2.10}$$

where

$$\Omega_{h_2}(x - x_t) = \int_{-\infty}^{x} W_{h_2}(u - x_t)\, du$$

The CDF estimator replaces the indicator function for a more standard estimation of equation (2.9) with a continuous distribution, Ω_{h_2}. The kernels have two different bandwidths, one each in the directions of x and y. The kernel W_{h_2} could be defined as being uniform, Gaussian or Epanechnikov. This approach combines both time and state domain smoothing of volatility.

While expression (2.10) provides a modeling framework for the CDF, it does not allow for quantile modeling along with regressors. A quantile regression framework incorporating the double kernel density approach has been presented by Taylor (2007):

$$\min_q \sum_{t=1}^{T} K_{h_1}(y - y_t) \left(\int_{\infty}^{\infty} (x - q)(\tau - \mathbb{1}(x < q))\, W_{h_2}(x - x_t)\, dx \right) \tag{2.11}$$

The above expression allows direct estimation of quantiles rather than iterative derivations from

CDF estimators in expression (2.10). The above minimization can be generalized to incorporate regressors, i.e. to estimate the quantile model $Q_t(\tau) = y_t' \beta$:

$$\min_{\beta} \sum_{t=1}^{T} K_{h_1}(y - y_t) \left(\int_{\infty}^{\infty} (x - y_t' \beta)(\tau - \mathbb{1}(x < y_t' \beta)) W_{h_2}(x - r_t) \, dx \right) \tag{2.12}$$

Furthermore, if we choose K_{h_1} to be the same exponentially weighted kernel considered before and choose W_{h_2} to be Gaussian, this minimization becomes:

$$\min_{\beta} \sum_{t=1}^{T} \lambda^{T-t} \left(\tau(x_t - y_t' \beta) + (y_t' \beta - x_t) \Phi((y_t' \beta - x_t)/h_2) + h_2 \phi((y_t' \beta - x_t)/h_2)\right), \tag{2.13}$$

where Φ and ϕ are the standard Gaussian CDF and probability density function respectively. The minimization problem can be solved using a nonlinear optimization algorithm. This approach is termed the exponentially weighted double kernel quantile regression (EWDKQR). To estimate the time-varying ES, the EWQR ES expressions (2.6) and (2.7) can be adapted in the double kernel framework. The EWDKQR ES expressions for quantiles in lower and upper tails are given by equations (2.14) and (2.15) respectively

$$\widehat{\mathrm{ES}}_T = -\frac{1}{\tau \sum_{t=1}^{T} \lambda^{T-t}} \sum_{t=1}^{T} \lambda^{T-t} \left(\tau(x_t - y_t' \hat{\beta}) + (y_t' \hat{\beta} - x_t) \Phi((y_t' \hat{\beta} - x_t)/h_2) + h_2 \phi((y_t' \hat{\beta} - x_t)/h_2)\right) \tag{2.14}$$

and

$$\widehat{\mathrm{ES}}_T = \frac{1}{(1-\tau) \sum_{t=1}^{T} \lambda^{T-t}} \sum_{t=1}^{T} \lambda^{T-t} \left(\tau(x_t - y_t' \hat{\beta}) + (y_t' \hat{\beta} - x_t) \Phi((y_t' \hat{\beta} - x_t)/h_2) + h_2 \phi((y_t' \hat{\beta} - x_t)/h_2)\right) \tag{2.15}$$

2.3 Modified EWDKQR method

We did a comparative study to compare different methods for estimating VaR and ES. We focused on two sets of methods for VaR estimation - the semi-parametric CAViaR method, and the non-parametric EWQR method in the single and double kernel framework. The results were comparable to those obtained by Taylor (2007) in their analysis across most of these estimation methods. It was interesting to see that the while the EWQR and EWDKQR (intercept only)

methods performed better than CAViaR, the double kernel method with regressors does not provide any added advantage over them. Further, since the regression parameter associated with these regressors was updated dynamically, there were often issues with convergence while solving the optimization problem. We proposed a modified way of implementing the double kernel approach with regressors that provided significant improvement over the original method. There were two key changes that we proposed - a) the step function that shifts the VaR forecast slightly based on the return of the previous day, and b) the optimization problem itself by treating the slope parameter as a global variable rather than updating it dynamically. The original approach uses an indicator variable that takes the value 1 if the log return, X_{t-1} was negative, and zero otherwise. The quantile estimate is then given as:

$$Q_t(\tau) \;=\; y_t'\beta \;=\; \beta_0 \;+\; \beta_1 \mathbb{1}_{\{x_{t-1} < 0\}} \tag{2.16}$$

This follows the rationale that we increase the VaR slightly if there is a downward movement in prices but leave it unchanged if there is an upward movement. We suggest replacing the indicator function by the signum function which suggests increasing the VaR in case of downward price movements, and decreasing in case of an upward change. The quantile estimate is then given as:

$$Q_t(\tau) \;=\; \beta_0 \;+\; \beta_1 \;\mathrm{sgn}(x_{t-1}) \tag{2.17}$$

where

$$\mathrm{sgn}(x) \;=\; \begin{cases} -1 & x \;<\; 0 \\ \;\;1 & x \;>\; 0 \end{cases}$$

Using a signum function helps produce forecasts that adapt quickly to the changes in the market as compared to using a indicator that provides uni-directional adjustment.

We solve the minimization problem in (2.12) using the training data to obtain the estimates for the kernel parameters λ, and the bandwidth h_2. This is thus a two-step optimization problem requiring us to estimate the regression coefficients at each time point for every iteration in the optimization for the kernel parameters. This makes the computations slow and tedious without a significant improvement in the accuracy of the estimates. We translate this double optimization problem to a single optimization by considering β_1 as a global variable, and estimating β_0 using the EWDKQR setup without regressors. A moving window of the past 250 observations is used to estimate the quantile at each time point. Since we move the estimation window by only a small amount, say one observation at each step, we do not lose much information by choosing a constant value of β_1. Further, the forecasts are more stable, and adapt quickly to the fluctuations in the series.

2.4 Empirical results

This section focuses on implementing the modified double kernel method in a comparative analysis framework with the other CAViaR and EWQR methods. Daily log returns for eight individual stocks from S&P 500 were considered, and their One-day ahead VaR forecasts were calculated . These are the same as considered by Taylor (2007) in his comparative study. The sample period considered is daily data for 13 years from 29 April 1992 to 29 April 2005. This resulted in 3279 observations of which the first 2779 returns are considered to estimate the parameters and the remaining 500 to evaluate daily one-step forecasts. The unconditional mean of the in-sample data is subtracted from the respective return series, and the quantile estimation procedure is then applied on the resulting residual series.

2.4.1 Methods used for VaR and ES estimation

1. EWQR and EWDKQR (intercept only model)
   For EWQR without regressors, a CDF estimator approach similar to expression (2.9) for the single kernel method was used to estimate the quantiles. While this approach is faster it cannot be used if there are regressors in the model. The quantiles in this case can be estimated by using linear programming to solve the minimization problem given in (2.8). An indicator variable was used to incorporate the "leverage" effect taking value 1 if the previous log return was negative, and 0 otherwise.

   Following the methodology in Taylor (2007), a moving window of 250 recent observations was used for each EWQR. Optimization of the λ was based on a rolling window of 250 observations to estimate the quantiles for the in-sample observations. The optimal value of λ is taken to be the one that minimises the quantile regression sum given in expression (2.3). While Taylor (2007) used a grid search in [0.8, 1] with a step size of 0.005, we used a global search in the same interval to get the optimal value.

   A Gaussian kernel was used to implement the double kernel method. Quantile estimation for EWDKQR without regressors was done using the CDF estimator similar to the single kernel approach. Values for the exponential scaling parameter λ in the interval [0.8, 1] and bandwidth h_2 in, (0, 0.02] were chosen subject to the minimization of the quantile regression sum. The starting value for the optimization process for λ was taken as the optimal value for λ from EWQR.

2. Modified EWDKQR Leverage Method
   The EWDKQR Leverage method was implemented similar to the EWDKQR (intercept only model) except for the addition of the indicator variable for negative log return at the

Table 2.1 Optimal slope values for Modified EWDKQR Leverage method

Quant	GE	XOM	MSFT	JNJ	PFE	WMT	INTC	PG
0.01	0.00310	0.00008	0.00081	0.00120	0.00121	0.00222	0.00164	-0.00111
0.05	0.00200	-0.00005	-0.00043	0.00039	0.00116	0.00038	0.00213	0.00074
0.95	0.00072	-0.00140	0.00082	-0.00089	-0.00006	-0.00066	-0.00175	-0.00179
0.99	0.00084	-0.00141	-0.00156	0.00137	0.00039	-0.00170	-0.00122	-0.00217
Based on 2779 in-sample observations								

previous time point. The optimal value for λ from the EWDKQR (intercept only) model was chosen as the starting value of the optimization of λ.

The modified EWDKQR Leverage method introduced by us replaces the indicator function with the signum function given in (2.17). Treating the beta coefficient associated with the signum function as a global parameter, we reduce the double optimization to a single optimization with one additional parameter. The optimal slope values at different quantiles are given in Table 2.1. The slope values for lower quantiles is mostly positive. A positive slope increases the quantile estimate or the VaR slightly when the previous return is negative i.e. there is an indication for downward movement in price. This ensures that the quantile estimate would adapt swiftly to the changing conditions.

3. CAViaR Models

   The CAViaR models were implemented here using a procedure described by Engle and Manganelli (2004). Their models are presented in the following expressions:

$$\text{Adaptive CAViaR} : Q_t(\tau) = Q_{t-1}(\tau) + \alpha_1(\tau - I(x_{t-1} < Q_{t-1}(\tau))),$$
$$\text{Symmetric Absolute Value CAViaR} : Q_t(\tau) = \omega + Q_{t-1}(\tau) + \beta|x_{t-1}|,$$
$$\text{Asymmetric Slope CAViaR} : \omega + \alpha_1 Q_{t-1}(\tau) + \beta_1(x_{t-1})^+ + \beta_2(x_{t-1})^-$$

   where $Q_t(\tau)$ is the conditional τ quantile; ω, α_1, β, and β_i are parameters; and $(u)^+ = max(u, 0)$, and $(u)^- = min(u, 0)$.

   For each model, we generated 10^5 vectors between 0 and 1 or -1 and 0 depending on the appropriate sign of the parameter using a uniform random generator. The quantile regression sum was evaluated for each of these vectors. The parameters corresponding to the 10 lowest values of the sum were taken as initial values for the optimization process to choose the final parameter vector.

The conditional quantile forecasts for postsample observations using the hit percentage and

the dynamic quantile (DQ) test described in section 1.5.2. The hit percentage assesses the unconditional coverage of a conditional quantile estimator at the τ level. The percentage should ideally be τ. A binomial distribution framework is used to assess significant differences. A DQ test statistic is used to evaluate the dynamic properties of the conditional quantile estimator. The test jointly evaluates that the hit variable, $\text{Hit}_t = \mathbb{1}_{\{x_t \leq \hat{Q}_t(\tau))\}} - \tau$, is distributed i.i.d. Bernoulli with probability τ, and is independent of the conditional quantile estimator. Following Engle and Manganelli's empirical study, four lags of Hit_t were included in the test's regression. The DQ test statistic under the null hypothesis is then distributed as a $\chi^2(6)$.

Table 2.2 presents the hit percentage or the percentage of VaR violations for all stocks series. The final column presents the number of stocks for which the hit percentage is significantly different from 5% at $\alpha = 5\%$ level of significance. The best results were obtained for Modified EWDKQR Leverage method followed by the EWQR and EWDKQR without regressors. Table 2.3 similarly presents the p-values for the DQ test. The best results were obtained with the EWQR Leverage and Adaptive (CAViaR) followed by EWDKQR without regressors and the Modified EWDKQR Leverage method. Tables 2.4 and 2.5 summarize the results of the Hit and DQ tests at 1%. The double kernel methods again perform really well. It should be noted that the performance of two CAViaR methods - symmetric absolute value (SAV) and asymmetric slope (AS) is comparable to the double kernel methods. However, the hit proportions or proportion of VaR exceedances in tables 2.2 and 2.4, and figure 4.1 show that there are very few VaR violations for these methods indicating that they may be too conservative.

Table 2.2 Evaluation of quantile forecasts at 5% level; hit proportion for 500 post-sample observations

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G	Number significant at 5% level
EWQR	5.8	9.6*	4.8	3.2	7	6.8	4.6	5.6	1
EWQR Leverage	3*	5.4	4.4	3*	4.6	4.8	3.2	6	2
EWDKQR	1.6*	5.6	3.4	3.6	5.4	3.2	4.6	5.4	1
Modified EWDKQR Leverage	4.8	5.8	3.8	3.8	5.6	4.4	4	6.6	0
Adaptive, CAViaR	4	4.6	2*	3.6	4.8	4.8	3.8	5.4	1
Symmetric Absolute Value, CAViaR	3.2	3.4	2.2*	1.8*	2.8*	1.2*	2.8*	1.8*	6
Asymmetric Slope, CAViaR	2*	3.6	2*	1.4*	2.6*	1.6*	3*	1.2*	7

* denotes significance at $\alpha = 5\%$

Table 2.3 Evaluation of quantile forecasts at 5% level; DQ test p-values for 500 post-sample observations

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G	Number significant at 5% level
EWQR	0.204	<0.0001*	0.850	0.556	0.001*	0.002*	0.465	0.300	3
EWQR Leverage	0.121	0.210	0.359	0.209	0.186	0.172	0.072	0.581	0
EWDKQR	0.007*	0.342	0.399	0.600	0.370	0.565	0.340	0.233	1
Modified EWDKQR Leverage	0.018*	0.555	0.904	0.619	0.322	0.199	0.169	0.009*	2
Adaptive, CAViaR	0.209	0.599	0.063	0.475	0.346	0.116	0.706	0.295	0
SAV, CAViaR	0.579	0.610	0.038*	0.092	0.130	0.020*	0.325	0.087	2
AS, CAViaR	0.140	0.718	0.038*	0.035*	0.081	0.057	0.475	0.020*	3

* denotes significance at $\alpha = 5\%$

Table 2.4 Evaluation of quantile forecasts at 1% level; hit proportion for 500 post-sample observations

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G	Number significant at 5% level
EWQR	3.2*	1.8	1	1	2.6*	0.6	4.2*	2.6*	4
EWQR Leverage	0.8	1.8	1.2	0.8	1.6	0.8	0.6	0.4	0
EWDKQR	0.4	1.6	0.4	0.4	0.8	0.6	0.4	0.6	0
Modified EWDKQR Leverage	0.6	2*	1.2	0.8	1.2	0.8	0.8	1.6	1
Adaptive, CAViaR	0*	0.6	0.4	0.6	0.6	0.6	0.8	1	1
Symmetric Absolute Value, CAViaR	0.2	0.4	0.4	0.2	1	0.6	0.4	0.4	0
Asymmetric Slope, CAViaR	0.2	0.8	0.4	0.2	1	0.4	0.4	0.4	0

* denotes significance at $\alpha = 5\%$

Table 2.5 Evaluation of quantile forecasts at 1% level; DQ test p-values for 500 post-sample observations

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G	Number significant at 5% level
EWQR	<0.0001*	0.242	0.765	0.9042	<0.0001*	0.8615	<0.0001*	0.0014	4
EWQR Leverage	0.9297	0.0259	0.9675	0.8343	0.0005*	0.8334	0.9649	0.9313	2
EWDKQR	0.9109	0.5678	0.9190	0.8383	0.0001*	0.8709	0.8413	0.9668	1
Modified EWDKQR Leverage	0.9171	0.0285*	0.6705	0.9913	0.0124*	0.9942	0.9753	0.4364	1
Adaptive, CAViaR	<0.0001*	0.8127	0.9137	0.7131	<0.0001*	0.8428	0.7192	0.3069	2
SAV, CAViaR	0.7633	0.9339	0.9140	0.7533	0.9930	0.9482	0.9200	0.9258	0
AS, CAViaR	0.7398	0.9912	0.9149	0.7530	0.9986	0.8750	0.9337	0.9224	0

* denotes significance at $\alpha = 5\%$

2.5 Summary

The EWQR method allows us to use exponential smoothing to estimate the time-varying quantiles of the conditional returns distribution. An added advantage of doing so in a regression framework is that it allows us to include market variables that help provide additional information for estimating these quantiles. The performance of the methods discusses here are competitive with the already established methods in the literature like the CAViaR methods of Engle and Manganelli (2004). The performance of the EWQR methods at the upper 1% suggests that it can be be considered a good candidate model for VaR estimation based on the Basel Committee recommendations. These methods should be further investigated for different datasets, and across varying market conditions to be able to comment on their robustness.

CHAPTER

3

VQR TEST: BACKTESTING VAR USING QUANTILE REGRESSIONS

3.1 Introduction

Section (1.5) discusses different strategies for backtesting VaR including the unconditional and conditional coverage tests by Kupiec (1995) and Christoffersen and Pelletier (2003) respectively, and the information variable based dynamic quantile (DQ) method by Engle and Manganelli (2004). These tests are based on an orthogonality condition between a binary variable (i.e, an indicator variable for whether the loss exceeded the VaR) and some information variables. The orthogonality condition holds if the VaR model is correctly specified. Berkowitz et al. (2011) showed that using this orthogonality condition, it is possible to interpret the existing backtests as Lagrange-Multiplier (LM)-type tests.

The LM-type tests usually have low power in finite samples against various misspecified models. The problem is especialliy critical to the VaR backtesting setup since binary variables are constructed to represent rare events. In finite samples where there are only few extreme events, the test may not have enough information to reject a misspecified model. This problem can be overcome by either increasing the sample size or constructing different tests. On these lines, Gaglianone et al. (2008) proposed using a random coefficient model to construct a Wald type test. They pointed out that LM and Wald tests may yield quite different results in small

samples even though they are asymptotically equivalent under the null hypothesis. The idea behind the test is to use extra information to reject the misspecified model thereby giving it more power in finite samples. The model and corresponding test statistic is discussed in detail in the next section.

3.2 Random coefficient model

It is clear from the mathematical definition of VaR in equation (1.1) that calculating VaR is equivalent to finding the conditional quantile of the series of portfolio returns, X_t. Gaglianone et al. (2008) adapted the idea of Christoffersen et al. (2001) that generating a VaR measure can be thought of as the outcome of a quantile regression with volatility as regressor to test the accuracy of a given VaR model by replacing the conditional volatility with the VaR measure of interest (V_t) as the regressor. Mathematically, a random coefficient model with V_t as regressor can be written as:

$$X_t = \beta_0(U_t) + \beta_1(U_t)V_t \tag{3.1}$$

$$= y_t'\beta(U_t) \tag{3.2}$$

where V_t is $\mathbb{F}_{t-1}$ measurable and already known at time $t-1$, $U(t) \sim$ iid $U(0,1)$, and $\beta_i(U_t),\ i = 0, 1,$ are assumed to be comonotonic in U_t, with $\beta(U_t) = [\beta_0(U_t), \beta_1(U_t)]'$ and $y_t' = [\mathbb{1}, V_t]$. Given the random coefficient model in equation (3.1) and the comonotonicity assumptions for the β coefficients, the conditional quantile of X_t at level τ can be written as:

$$Q_{X_t}(\tau|\mathbb{F}_{t-1}) = \beta_0(\tau) + \beta_1(\tau)V_t \quad \text{for all } \tau \in (0,1). \tag{3.3}$$

Considering the quantile model in equation (3.3), a test to assess the overall performance of VaR can be constructed to by testing the following null hypothesis:

$$H_0 : \begin{cases} \beta_0(\tau) = 0 \\ \beta_1(\tau) = 1 \end{cases} \tag{3.4}$$

against the general alternative. The null hypothesis can also be written in a classical formulation as $H_0 : \theta(\tau) = 0$ where $\theta(\tau) = [\beta_0(\tau),\ \beta_1(\tau) - 1]$.

Let $\hat{\theta}(\tau)$ be the quantile regression estimator of $\theta(\tau)$. The asymptotic distribution of $\hat{\theta}(\tau)$ is normal with covariance matrix that takes the form of a Huber (1967) sandwich:

$$\sqrt{T}(\hat{\theta}(\tau) - \theta(\tau)) \xrightarrow{d} N(0,\ \tau(1-\tau)H_\tau^{-1}JH_\tau^{-1}) \tag{3.5}$$

where $J = \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} y_t y_t'$ and $H_\tau = \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} y_t y_t' [f_t(Q_{X_t}(\tau|y_t))]$ under the quantile regression model $Q_{X_t} = y_t'\beta(\tau)$ where $\theta(\tau) = \beta(\tau) + [0, -1]'$ and $f_t(Q_{X_t}(\tau|y_t))$ represents the conditional density of the return X_t at the quantile τ. Given that we are able to compute consistent estimators of H_τ and J, the test statistic for testing the null hypothesis in (3.4) can be defined as:

$$\text{VQR} = T[\hat{\theta}(\tau)' \, (\tau(1-\tau) \, H_\tau^{-1} \, J \, H_\tau^{-1})^{-1} \, \hat{\theta}(\tau)] \tag{3.6}$$

In addition, under the null hypothesis in (3.4), if the assumptions A1 through A4 (given below) hold, the test statistic in (3.6) is asymptotically chi-squared distributed with 2 degrees of freedom.

A1 Y_t is $\mathbb{F}_{t-1}$ measurable and $z_t \equiv \{X_t, Y_t\}$ is a strictly stationary process.

A2 X_t has conditional (on y_t) distribution functions F_t, with continuous Lebesgue densities f_t uniformly bounded away from 0 and ∞ at the points $Q_{X_t}(\tau|y_t) = F_t^{-1}(\tau|y_t)$ for all $\tau \in (0, 1)$.

A3 There exist positive density matrices J and H_τ, such that for all $\tau \in (0, 1)$:

$$J = \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} y_t y_t' \tag{3.7}$$

$$H_\tau = \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} y_t y_t' [f_t(Q_{X_t}(\tau|y_t))] \tag{3.8}$$

A4 $\max_{t=1,\ldots,T} ||y_t|| / \sqrt{T} \xrightarrow{p} 0$.

3.2.1 Covariance matrix estimation

A key step in computing the covariance matrix is the estimation of the H_τ matrix and subsequently the estimation of the conditional quantile density at quantile τ. This density can be estimated as the reciprocal of the "sparsity function" at the quantile τ, $s(\tau)$. Mathematically,

$$s(\tau) = [f(F^{-1}(\tau))]^{-1}$$

The asymptotic precision of the quantile estimates in quantile regression depends on the sparsity function since it reflects the density of observations near the quantile of interest. If the sparsity is high, there are few observations near the quantile of interest and hence it would be difficult to estimate. It is also easy to show that the sparsity function is the derivative of the quantile

function:

$$F(F^{-1}(u)) = u \implies \frac{d}{du}F^{-1}(u) = s(u)$$

Siddiqui (1960) estimated the sparsity function using a simple difference quotient of the empirical quantile function:

$$\hat{s}_T(\tau) = \frac{\hat{F}_T^{-1}(\tau + h_T) - \hat{F}_T^{-1}(\tau - h_T)}{2h_T} \tag{3.9}$$

where $\hat{F}_T^{-1}$ is an estimate of F^{-1} and h_T is a bandwidth that tends to zero as $T \to \infty$. Hendricks and Koenker (1991) adapted the sparsity function in (3.9) into the quantile regression framework and showed that if the τth conditional quantile function of $x|y$ is linear, then as $h_T \to 0$, the density can be estimated by the difference quotient

$$\hat{f}_t(Q_{X_t}(\tau|y_t)) = \frac{2h_T}{y_t'(\tau + h_T)\ \hat{\beta}(\tau + h_T) - y_t'(\tau - h_T)\hat{\beta}(\tau - h_T)} \tag{3.10}$$

where $y_t(\tau \pm h_T) \ = \ [1, V_t(\tau \pm h_T)]$, and $V_t(\tau \pm h_T)$ is the VaR evaluated at points $\tau \pm h_T$.

Since there is no guarantee for the proposed density estimate to be positive for every observation in the sample, it can be replaced by

$$d_t = y_t'(\tau + h_T)\ \hat{\beta}(\tau + h_T) - y_t'(\tau - h_T)\hat{\beta}(\tau - h_T) \tag{3.11}$$

$$\hat{f}_t^{+}(Q_{X_t}(\tau|y_t)) = \max\{0,\ 2h_T/(d_t - \epsilon)\}, \tag{3.12}$$

where $\epsilon > 0$ is a small tolerance parameter to avoid dividing by zero. A simpler kernel based technique can also be used to estimate the density at the required quantile as given in Powell et al. (1991).

3.2.2 Choice of bandwidth h_T

Bofinger (1975) used the standard density estimation asymptotics to show that

$$h_T \ = \ T^{-1/5}\ [4.5\ s^2(\tau)/(s''(\tau))^2]^{1/5} \tag{3.13}$$

minimises the means squared error. If both $s(\tau)$ and $s''(\tau)$ are known, it is not required to estimate h_T but since $s(\tau)/s''(\tau)$ is not very sensitive to F, it can be substituted with typical distributional shapes without losing much information. The ratio

$$\frac{s}{s''} \ = \ \frac{f^2}{2(f'/f)^2 \ + \ [f'/f \ - \ f''/f]}$$

is scale and location invariant. If f is chosen as the standard normal density, i.e. $f \ = \ \phi$,

$(f'/f)(F^{-1}(t)) \;=\; \Phi^{-1}(t)$, and

$$h_T \;=\; T^{-1/5} \left[\frac{4.5\;\phi^4\;(\Phi^{-1}(t))}{2\Phi^{-1}(t)^2 \;+\; 1)^2}\right]^2 \tag{3.14}$$

Another choice of the bandwidth sequence was suggested by Hall and Sheather (1988) based on Edgeworth expansions for studentized univariate sample quantiles where

$$h_T \;=\; T^{-1/3}\; z_\alpha^{2/3}\; [1.5\; s(\tau)/s''(\tau)]^{1/3} \tag{3.15}$$

where z_α satisfies $\Phi(z_\alpha) \;=\; 1 - \alpha/2$, and α is the desired size of the test. These bandwidths are narrower for moderate to large t compared to those suggested by Bofinger (1975).

3.3 Simulation study

We conduct a simulation experiment to compare the VQR test with the unconditional coverage test by Kupiec (1995) and dynamic quantile (DQ) test by Engle and Manganelli (2004) by evaluating the empirical size of 5% tests. We simulate 5000 sample paths of length T where $T \;=\; \{250, 500, 1000, 2500\}$ i.e. approximately 1, 2, 4, and 10 years of daily data. The correct data generating process is assumed to be a zero mean, unit unconditional variance GARCH model with Gaussian innovations:

$$R_t \;=\; \sigma_t \epsilon_t,\; t \;=\; 1, \ldots, T, \tag{3.16}$$

$$\sigma_t^2 \;=\; (1 \;-\; \alpha \;-\; \beta) \;+\; \alpha\, y_{t-1}^2 \;+\; \beta\, \sigma_{t-1}^2 \tag{3.17}$$

where $\epsilon_t \;\sim\; iid\; N(0,1)$. We consider 2 different models based on the values of α and β:

- Model 1: $\alpha \;=\; 0.05\;,\; \beta \;=\; 0.9$
- Model 2: $\alpha \;=\; 0.5\;,\; \beta \;=\; 0.45$

The GARCH reaction parameter, α, for daily data usually ranges between 0.05 (a relatively stable market) to 0.1 (a jumpy or nervous market). The persistence parameter, β, usually ranges between 0.85 to 0.98, with lower values associated with higher α. A model with high α and a low β suggests high GARCH volatilities with high vol-of-vol. The true VaR can be computed as

$$V_t \;=\; \sigma_t\; \Phi_\tau^{-1} \tag{3.18}$$

where Φ_τ^{-1} denotes the τ quantile of a standard normal distribution. Since we are computing 5% tests, we would expect that each test would reject the correctly specified model (3.18) 5% of times under the null hypothesis.

Table 3.1 Empirical Size of 5% Tests, $\tau = 0.95$

	Model 1 $\alpha = 0.05$, $\beta = 0.9$				Model 2 $\alpha = 0.5$, $\beta = 0.45$			
Sample Size/Method	250	500	1000	2500	250	500	1000	2500
Hit	0.0404	0.0354	0.0476	0.0508	0.0404	0.0356	0.0476	0.0524
DQ1	0.045	0.042	0.043	0.044	0.037	0.036	0.031	0.039
DQ5	0.0722	0.061	0.0544	0.0562	0.062	0.056	0.0526	0.0556
VQR	0.0734	0.0572	0.054	0.0542	0.0824	0.0636	0.0584	0.0524

Table 3.1 reports the size of tests for $\tau = 0.95$. Two dynamic quantile (DQ) tests, DQ1 and DQ5 depending on the choice of instrument variables. DQ1 is analogous to the VQR test with instrument variable $X_t = [1, V_t]$ whereas DQ5 builds on DQ1 by using four lagged values of VaR in addition to X_t. The size of the VQR test is calculated using asymptotic critical values from $\chi^2(2)$ distribution.

The bandwidth sequences described in (3.13) and (3.15) assume a Gaussian distribution and thus give narrow windows for extreme quantiles. A modification of the bandwidth sequence in (3.15) is considered here:

$$h_{T*} = \frac{4}{3} T^{-1/3} z_\alpha^{2/3} [1.5\, s(\tau)/s''(\tau)]^{1/3} \tag{3.19}$$

The modified bandwidth calculation allows us a wider window to estimate the conditional density in (3.10) and yields a stable sandwich estimate for the VQR test statistic in (3.6). However, the technique cannot be used for more extreme quantiles say $\tau = 1\%$.

It is evident from tables that the asymptotic critical values do not work well when sample sizes are small. The VQR test responds to it by rejecting the null model more frequently than 5% of the time. However, if we allow for large sample sizes, the empirical size of the test converges to the nominal size at the 95% level. The unconditional coverage test seems to be undersized for smaller sample sizes but converges to the nominal size as the sample increases. DQ tests however are oversized for smaller samples but converge as the sample increases. The observations are dense at the 95% quantile level and consequently help in accurately estimating the conditional quantiles.

While modifying the bandwidth sequence helped in the previous case, it is not easily applicable at the 99% level. We therefore use the standard approach in Hall and Sheather (1988) to estimate the conditional density . Since the bandwidth is extremely small, there is an evident issue of quantile crossing in (3.11) which leads the estimated conditional density to be zero at a significant number of points. A modified 2-step approach is considered here:

Table 3.2 Empirical Size of 5% Tests, $\tau = 0.99$

	Model 1 $\alpha = 0.05$, $\beta = 0.9$				Model 2 $\alpha = 0.5$, $\beta = 0.45$			
Sample Size/Method	250	500	1000	2500	250	500	1000	2500
Hit	0.0404	0.0354	0.0476	0.0508	0.0404	0.0356	0.0476	0.0524
DQ1	0.0484	0.0460	0.0454	0.0494	0.0324	0.0344	0.0322	0.0354
DQ5	0.1790	0.1766	0.1052	0.1094	0.176	0.1734	0.1088	0.1048
VQR	0.2742	0.1750	0.1102	0.0474	0.1026	0.0876	0.0644	0.0510

- Step 1: Since the bandwidth h_T is very small for $\tau = 0.99$, we replace the quantile regression coefficients estimated for $\tau \pm h_T$ to be the same as that for level τ. The resulting density estimate therefore is

$$\hat{f}_t^+(Q_{X_t}(\tau|y_t)) = max\left\{0,\ \frac{2h_T}{\hat{\beta}(\tau)\ [y_t'(\tau + h_T)\ -\ y_t'(\tau - h_T)] - \epsilon}\right\} \tag{3.20}$$

- Step 2: Points with estimated density zero in Step 1 are calculated using the original formula in (3.11).

$$\hat{f}_t^+(Q_{X_t}(\tau|y_t)) = max\left\{0,\ \frac{2h_T}{\hat{\beta}(\tau)\ [y_t'(\tau + h_T)\ \hat{\beta}(\tau + h_T) - y_t'(\tau - h_T)\hat{\beta}(\tau - h_T)] - \epsilon}\right\} \tag{3.21}$$

The order of the steps ensures that the density estimate is calculated with more accuracy and has fewer points where it is zero. Table 3.2 gives the empirical sizes for $\tau = 99\%$. It can be seen that the unconditional coverage test which is an exact binomial test for VaR exceedances is undersized in both scenarios. The DQ1 test is undersized as well but performs much better in a stable market situation. The DQ5 test, which has more information from the lagged values of VaR, is however oversized and does not converge to the nominal size of the test even for larger samples. This implies that the tests are rejecting the null model less often than they should be.

The VQR test is significantly oversized for smaller sample sizes but converges to the nominal size of the test as sample size increases across the two models. It should also be noted that the performance of the VQR test is better in Model 2 which has a high vol-of-vol as compared to Model 1. The variation of volatility is important for the performance of the quantile regression since it needs some range in the predictor variables for better estimation of the parameters. While the DQ1 and VQR tests both have the same instrument variable $Y_t = [1, V_t]$ with V_t being the VaR at quantile level τ, they test different things based on their setup. The DQ test uses a VaR violation indicator as response in a least squares regression setup whereas the VQR test uses the conditional quantile of the return series at level τ in a quantile regression framework.

Backtests that use binary responses like the DQ test are based on orthogonality conditions based on the orthogonality condition between the binary variables (here, VaR violation indicator) and other instrument variables. An LM test can be derived here if the VaR model is specified correctly and consequently the orthogonality condition holds. These tests have been shown to have low power in case of finite samples when there are few extreme events and the test does not have the information needed to reject a misspecified model. It may be possible to deal with this by increasing the sample size. The VQR test provides a different approach for evaluating the null hypothesis that the VaR model is correctly specified. Even though LM and Wald tests are asymptotically equivalent, the simulation studies here show that they may yield different results for finite samples. The empirical sizes at 5% for the DQ1 and VQR tests in table 3.1 show that the DQ1 does not reject the null model as often as expected even for larger sample sizes. The difference in empirical sizes here may be attributed to the additional information available in the VQR setup.

The difference between these tests can also be investigated with respect to their power to reject a misspecified model. Two misspecified alternatives for X_t are considered here:
H_{A_1}: GARCH model with underestimated conditional variances: $\sigma^2_{t,\kappa} = (1 - \kappa) \times \sigma^2_t$ where $\kappa = 0.25$.
H_{A_2}: 12-month historical simulation model: VaR is calculated empirically from the 250 previous trading days such that $VaR_t(\tau)$ is the τ-th percentile of X_t.

In H_{A_1}, the conditional variance of the series σ_t is underestimated by 25% in the alternative hypothesis to examine if the tests are able to detect an underestimation in the VaR occurring due to a misleading appreciation of volatility. In H_{A_2}, the misspecification occurs because the historical simulation method is unable to correctly capture the time-varying dynamics in the return series. The alternatives chosen here are in line with the existing literature on risk assessment by Gaglianone et al. (2008), Couperier and Leymarie (2019), among many others.

Table 3.3 reports the size adjusted power of 5% test for Models 1 and 2. For H_{A_1}, the power of all the tests across different sample sizes is high for both the models. This indicates that all the models were able to detect the underestimation in conditional density under the alternative hypothesis in either market scenario. The response of the power analysis under H_{A_2} however is starkly distinct across the two models. For Model 1, the size adjusted power is low, and often smaller than the nominal size for the Hit, DQ1, and VQR tests. The low power for the Hit, and DQ1 test can be attributed to fairly similar VaR estimates using the historical simulation, and the parametric VaR under the GARCH model in a stable market situation. The VQR test suffers additionally from the poor performance of quantile regressions in this case since the moving window is large enough to yield rather constant estimates of VaR.

In Model 2, the size-adjusted power for H_{A_2} is slightly higher than the power in Model 1. Given that the model characterizes a high vol-of-vol scenario, it is expected that the historical

Table 3.3 Size-adjusted Power of 5% Tests

		Model 1 $\alpha = 0.05$, $\beta = 0.9$				Model 2 $\alpha = 0.5$, $\beta = 0.45$			
H_a	Sample Size/Method	250	500	1000	2500	250	500	1000	2500
Scaled Volatility	Hit	0.9646	0.9988	1	1	0.9646	0.9988	1	1
	DQ5	0.8634	0.991	0.9998	1	0.8684	0.9984	1	1
	DQ1	0.9312	0.9968	1	0.9998	0.942	0.9914	1	1
	VQR	0.3696	0.9336	1	1	0.3018	0.923	1	1
Historical Simulation	Hit	0.1174	0.0478	0.0106	0.0056	0.2696	0.1658	0.0812	0.1086
	DQ5	0.1256	0.0652	0.04	0.0474	0.6638	0.9052	0.9916	1
	DQ1	0.1418	0.1564	0.1822	0.307	0.3792	0.34	0.3434	0.5614
	VQR	0.0782	0.0398	0.0302	0.0354	0.1482	0.171	0.2194	0.4432

simulation method would not be able to adapt to the dynamic changes in the dataset, and the tests therefore should be able to reject it. For the same set of information variables, the DQ1 test has higher power than the VQR test across all sample sizes. The DQ5 test considers VaR violation indicators at past four lags in addition to VaR as regressors. The DQ5 test benefits from the additional information through the lagged VaR violation indicators, and has the highest power of the methods considered across the two models.

3.4 Empirical analysis

As in Section 2.4, VQR test is used to compare the performance of VaR estimates calculated with different methods. Daily log returns for 8 individual stocks from S&P 500 were considered. The sample period considered is daily data for 13 years from 29 April 1992 to 29 April 2005. This resulted in 3279 observations of which the first 2779 returns are considered to estimate the parameters and the remaining 500 to evaluate daily one-step forecasts.

It is clear from tables 3.1 and 3.2 that asymptotic critical values do not do a good job for smaller sample sizes. One way to take into account the effect of size distortions is to calculate the Monte Carlo p-values instead. Each of these stock series was fitted using a GARCH model with normal-innovations and the fitted model was used calculate the p-values. These p-values are based on 5000 simulations and presented in Table 3.4. The last column in the table tallies the number of significant values for each method where significance implies that the null model was rejected.

The results in table 3.4 are similar to those in tables (2.2) and (2.3) in that the exponential

Table 3.4 Evaluation of quantile forecasts at 5% level;VQR test for 500 post-sample observations

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G	# of significant values
EWQR	0.0822	0.1438	0.057	<0.0001*	0.0536	0.0052*	0.0032*	0.387	3
EWQR Leverage	0.1116	0.2280	0.0188*	0.0576	0.0582	0.0064*	0.0086*	0.3690	3
EWDKQR	0.0154*	0.4602	0.0912	0.0180*	0.7504	0.0094*	0.0080*	0.0058*	5
EWDKQR Leverage	0.0210*	0.7110	0.0588	0.0026*	0.5884	0.006*	0.004*	0.2130	4
Adaptive, CAViaR	0.1114	0.0002*	0.0082*	0.0124*	0.4020	0.0292*	0.01060	0.9916	5
SAV, CAViaR	0.2052	0.3722	0.0082*	0.0474*	0.0140	0.0052*	0.2886	0.0052*	5
AS, CAViaR	0.0106*	0.0998	0.0078*	0.0110*	0.0150*	0.0034*	0.1046	0.0044*	6

* denotes significance at 5% level

weighted methods perform better than the CaViaR methods. Table 3.5 summarizes the performance of the three tests with respect to the frequency with which the null model was rejected. The relative performance of these methods for these stocks is similar to the observations made in the simulation studies. The Hit and DQ tests (lag=4) seem to be less conservative than the VQR test. Moreover, the modified EWDKQR technique performs well consistently when compared with the three techniques.

Table 3.5 Comparison of VQR test with Hit and DQ tests

	Number of significant values		
	Hit Test	DQ Test	VQR Test
EWQR	1	3	3
EWQR Leverage	2	0	3
EWDKQR	1	1	5
EWDKQR Leverage	0	2	4
Adaptive, CAViaR	1	0	5
SAV, CAViaR	6	2	5
AS, CAViaR	7	3	6

3.5 Summary

In this chapter, we discuss the use of a quantile regression based backtest for VaR and outline its implementation for risk levels in the extreme upper tail. Most VaR backtests investigate the VaR accuracy by constructing tests based on hit indicators that signal whether a particular risk threshold was exceeded. Using quantile tests like the VQR take into account the magnitude of

realized profits and losses and thus provide additional power to detect an inaccurate risk model. This is also reflected in the consistent performance of the VQR backtests across different market scenarios.

Since the VaR model can be implicitly defined as a conditional quantile function, it is possible to construct a backtest fully based on the quantile regression framework. A Wald type hypothesis test can then be used to evaluate the performance of VaRs. An important drawback of using these tests is that the underlying asymptotics do not hold well for extreme quantiles or in this case risk levels in the far upper tail. The objective of this chapter is to provide a working guideline to use this framework to backtest extreme quantiles. The proposed methodology allows for a larger bandwidth h_T for estimating the Hendricks-Koenker sandwich matrix, and consequently a more stable estimate using the added information. The modification performs well at both $\tau = 5\%$ and $\tau = 1\%$ but quantile crossing is more persistent at $\tau = 1\%$ and requires special consideration. The 2-step approach combining the original and proposed methodologies provides significant improvement to the estimation of the conditional density, and consequently the backtest. While the modification serves well for 1%, a possible alternative is to use a constrained quantile regression approach to avoid quantile crossing at the extreme quantiles.

An advantage of using the VQR test is that since the backtest does not depend on the underlying VaR, it is applicable to a wide variety of structures. While the backtest summarizes the global performance of the VaR model, it can also be extended to identify the periods in which the VaR model might increase the risk exposure, as outlined in Gaglianone et al. (2008). One of the benefits of using VQR was expected to be the increased power of the test in comparison to the other established backtests. The Monte Carlo studies presented here do not confirm to the expectations. Finally, while the method warrants further investigation, it is still an interesting prospect for its conceptual simplicity, and computational ease using standard quantile regression softwares.

CHAPTER

4

A COMPARATIVE EVALUATION OF TECHNIQUES FOR BACKTESTING EXPECTED SHORTFALL

4.1 Introduction

Expected shortfall is the conditional expected loss given a VaR exceedance at a given level τ. It is a coherent risk measure and a more sensitive one compared to VaR. ES, for this reason, has been getting increased attention from regulatory authorities in the recent years. However, the backtesting literature on ES is relatively small. There are several non-parametric approaches for backtesting expected shortfall. We group them based on procedural similarities into two categories- assuming asymptotic distribution, and quantile approximation. These methods are discussed in detail in the upcoming sections.

4.2 Asymptotic distribution

The first approach based on a normally distributed Z-test was proposed by Costanzino and Curran (2015). They propose a backtest for any spectral risk measure using the Central Limit Theorem (CLT) to approximate the normal distribution. Every measure in the class of spectral

risk measures is coherent. Costanzino and Curran (2015) show that ES can be viewed as a spectral risk measure and thus can be backtested in this general way.

Definition 4.2.1. A risk spectrum $\phi \in L^1([0,1])$ is said to be admissible if ϕ is non-negative, non-increasing, and $||\phi||_1 = 1$.

Definition 4.2.2. Let X be a random variable with cumulative distribution function F_X and ϕ an admissible risk spectrum. A spectral risk measure, M_ϕ with risk spectrum ϕ is defined by

$$M_\phi = \int_0^1 \phi(u)\,\mathrm{VaR}(u)\,du. \tag{4.1}$$

M_ϕ depends on the distribution of X through the VaR. If $\phi(u) = \mathrm{Dirac}_\alpha(u)$, then $M_{\mathrm{Dirac}_\alpha} = \mathrm{VaR}(\alpha)$. However, VaR is not a spectral risk measure because $\mathrm{Dirac}_\alpha(u)$ is not admissible as it violates properties i and ii in Definition 4.2.1.

Definition 4.2.3. For an admissible spectrum ϕ, $\psi^{(i)}_{\phi_\tau} \in [0,1]$, the failure indicator is defined as

$$\psi^t_{\phi_\tau}(X) = \int_0^1 \phi(u)\mathbb{1}_{\{X_t > VaR_t(u)\}}\,du. \tag{4.2}$$

The spectral risk measure failure rate $\Psi^T_{\phi_\tau} \in [0,1]$ is then defined as

$$\Psi^T_{\phi_\tau}(X) = \frac{1}{T}\sum_{t=1}^{T}\psi^t_{\phi_\tau}(X) \tag{4.3}$$

The VaR violations are discrete events with value either zero or one but failures corresponding to the spectral measures are continuous variables in $[0,1]$. The null hypothesis for the spectral risk measure coverage test is

$$H_0 : \{\psi^t_{\phi_\tau}(X)\}_{t=1}^T \text{ are i.i.d, and } P[X_t \geq \mathrm{VaR}_t(u)] = u \ \forall\, u \in \text{ supp } \phi \tag{4.4}$$

Expected shortfall is a special case of a spectral risk measure with risk spectrum given by

$$\phi_{ES} = \frac{1}{1-\tau}\mathbb{1}_{\{\tau \leq u \leq 1\}} \tag{4.5}$$

and the ES failure rate then becomes

$$\Psi^T_{ES_\tau}(X) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{1-\tau}\int_\tau^1 \mathbb{1}_{\{X_t \geq \mathrm{VaR}(u)\}}\,du \tag{4.6}$$

The expected shortfall measure M_{ES} admits a Z-test with Z-score

$$Z^T_{\text{ES}}(\tau) = \sqrt{3T}\,\frac{2\widehat{\Psi}^T_{ES_\tau}(X) - (1-\tau)}{\sqrt{(1-\tau)(4-3(1-\tau))}} \tag{4.7}$$

The most intuitive explanation of this backtest is that it calculates the probability of cumulative VaR exceedances in a sample. The backtest though simple in ease of application does not inform which part of the null hypothesis in expression (4.4) is rejected. The test should be further supplemented with a test for autocorrelation to make conclusions. This backtest can thus be expected to need more VaR exceedances to reject the hypothesis.

4.3 Quantile approximation

The third methodology based on approximating ES by VaR at different confidence levels was proposed by Emmer et al. (2015). This method even though a rough approximation is not as complex as the other techniques available in the literature. ES can be represented as integrated VaR

$$\text{ES}_\tau(X) = \frac{1}{1-\tau}\int_\tau^1 \text{VaR}_\tau(X)\,d\iota \tag{4.8}$$

$$\approx \frac{1}{4}\left[\text{VaR}_\tau(X) + \text{VaR}_{0.75\tau+0.25}(X) + \text{VaR}_{0.5\tau+0.5}(X) + \text{VaR}_{0.25\tau+0.75}(X)\right] \tag{4.9}$$

These component VaR terms are elicitable and can be backtested using existing methodologies. If all these levels are successfully backtested, ES can be considered reliable if the exceedances are actually far tail observations and not data outliers. This according to Emmer et al. (2015) provides a combination of testing and human oversight for backtesting expected shortfall.

Emmer et al. (2015) also observe that the power of the joint test for VaR violations on the chosen quantiles will decline with the number of quantiles considered but increase with the size of the sample observations. The choice of the number of quantiles considered is thus subjective and should be decided on a case by case basis. It implies that a much longer sample is needed for validation of ES than for VaR at the same confidence level.

4.3.1 Method 2: Independent testing

In order to backtest ES, VaR at these levels should be tested jointly. These levels can be tested either independently or tested jointly for unconditional coverage. It does not make sense to test these independently since VaR violations at different levels would not be independent. The

dependence implies that a lot of methods that combine multiple tests cannot be used. The correlation matrix based on sample size 500 and 6 levels of VaR for $\tau = 0.95$ for model 2 in Section 3.3 is:

$$\begin{pmatrix} 1.000 & 0.972 & 0.915 & 0.755 & 0.506 & 0.391 \\ 0.972 & 1.000 & 0.941 & 0.776 & 0.520 & 0.402 \\ 0.915 & 0.941 & 1.000 & 0.825 & 0.553 & 0.427 \\ 0.755 & 0.776 & 0.825 & 1.000 & 0.670 & 0.518 \\ 0.506 & 0.520 & 0.553 & 0.670 & 1.000 & 0.773 \\ 0.391 & 0.402 & 0.427 & 0.518 & 0.773 & 1.000 \end{pmatrix}$$

It is clear from the above matrix that the correlation is higher between VaR violations at levels closer to each other which is as expected. If these tests are assumed to be independent Fisher's method to combine the p-values from the tests as:

$$\chi^2_{2k} \;=\; -2 \sum_{k=1}^{p} ln(p_k) \tag{4.10}$$

where k levels are considered, and p_k is the p-value for the k^{th} hypothesis test. Another alternative is to consider the minimum p-value of the individual tests as the test statistic.

4.3.2 Method 3: Joint unconditional coverage

In Section 1.5.1.1, Hit test by Kupiec (1995) was used to test the unconditional coverage for VaR at a given level using a binomial null distribution. The test has been extended by Kratz et al. (2016) to jointly test VaR at multiple levels using a multinomial null distribution. The idea can be generalized by considering the VaR probability levels $\tau_1, \ldots, \tau_n$ defined as

$$\tau_j \;=\; \tau \;+\; \frac{j-1}{n}\,(1 \;-\; \tau), \qquad j = 1, \ldots, n, \;\; n \in \mathbb{N} \tag{4.11}$$

The value of τ is usually set to 0.975 corresponding to the level used for expected shortfall calculation. Let $I_{t,j} \;= \mathbb{1}(X_t > VaR_{\tau_j,t})$ be the VaR violation or exceedance indicator at level τ_j at time t and $Z_t \;=\; \sum_{j=1}^{n} I_{t,j}$. Then the the sequence $(Z_t)_{t=1,2,\ldots,T}$ counts the number of VaR levels that are breached. The sequence (Z_t) should satisfy two conditions:

- the *unconditional coverage* hypothesis, $P(Z_t \;\leq\; j) \;=\; \tau_{j+1},\; j = 0, \ldots, n$ for all t
- the *independence hypothesis*, Z_t is independent of Z_s for $s \neq t$.

The unconditional coverage probability can alternately be written as

$$Z_t \;\sim\; MN(1, (\tau_1 - \tau_0), \ldots, (\tau_{n+1} - \tau_n)), \;\; \text{for all } t$$

where $\tau_0 = 0$, $\tau_{n+1} = 1$ Let the observed cell counts be denoted by O_j, then

$$O_j \;=\; \sum_{t=1}^{T} I_{\{Z_t=j\}}$$

and the random vector $(O_0, \ldots, O_n)$ should follow a multinomial distribution

$$(O_0, \ldots, O_n) \;\sim\; MN(T, (\tau_1 - \tau_0), \ldots, (\tau_{n+1} - \tau_n))$$

To formalize the hypothesis testing problem, let $0 = \theta_0 < \theta_1 < \ldots < \theta_n < \theta_{n+1} = 1$ be an arbitrary sequence of parameters and consider the model where $(O_0, \ldots, O_n) \;\sim\; MN(T, (\theta_1 - \theta_0), \ldots, (\theta_{n+1} - \theta_n))$. Then

$$\begin{aligned} H_0 &: \theta_j \;=\; \tau_j \;\text{ for } j = 1, \ldots, n \\ H_1 &: \theta_j \;\neq\; \tau_j \;\text{ for at least one } j \;\in\; \{1, \ldots, n\}. \end{aligned} \tag{4.12}$$

Various tests can be used to evaluate these hypotheses of multinomial proportions including and not limited to a likelihood ratio test (LRT), the standard Pearson chi-square test, and the Nass test. Since the observed cell counts tally the VaR violations out in the tail, it is possible that the cell sizes are extremely small. The Nass test in this case is said to offer an appreciable improvement over the chi-square test.

- Pearson chi-squared test: The test statistic is

  $$S_n \;=\; \sum_{j=0}^{n} \frac{(O_{j+1} \;-\; n(\tau_{j+1} \;-\; \tau_j))^2}{n(\tau_{j+1} \;-\; \tau_j)} \;\underset{H_0}{\overset{\text{d}}{\sim}}\; \chi^2_n$$

  and the null hypothesis is rejected when $S_n > \chi^2_n(1-\alpha)$ where $\chi^2_n(1-\alpha)$ is the $(1-\alpha)-$quantile of the χ^2_n-distribution.

- Nass test: An improved approximation of the distribution of S_n was suggested by Nass defined as

  $$cS_n \;\underset{H_0}{\overset{\text{d}}{\sim}}\; \chi^2_\nu, \qquad \text{with } c \;=\; \frac{2E(S_n)}{var(S_n)} \text{ and } \nu \;=\; cE(S_n),$$

  where $E(S_n) = n$ and $\text{var}(S_n) = 2n - \frac{n^2+4n+1}{T} + \frac{1}{T}\sum_{j=0}^{n} \frac{1}{\tau_{j+1}-\tau_j}$. The null distribution is rejected when $cS_n > \chi^2_\nu(1-\alpha)$.

4.3.3 Method 4: Multi-quantile regression

Chapter 3 outlines the use of quantile regression to backtest VaR at a given probability level. This method has been further extended by Couperier and Leymarie (2019) to backtest expected

shortfall using multiple quantiles. ES at probability level τ may be approximated as the Riemann sum of VaR at n probability levels given in 4.11, i.e.

$$ES_t(\tau) \equiv \frac{1}{n} \sum_{j=1}^{n} \text{VaR}_t(\tau_j) \tag{4.13}$$

Extending the approach in Chapter 3, the ex-post losses $\{X_t, t = 1, 2, \ldots, T\}$ are regressed on the n VaR forecasts $\{\text{VaR}_t(\tau_j), t = 1, 2, \ldots, T\}_{j=1,2,\ldots,n}$ in a multi-quantile regression model:

$$X_t = \beta_0(\tau_j) + \beta_1(\tau_j)\, \text{VaR}_t(\tau_j) + \epsilon_{j,t} \quad \forall j = 1, 2, \ldots, n \tag{4.14}$$

where $\beta_0(\tau_j)$, and $\beta_1(\tau_j)$, denote the intercept and slope respectively at level τ_j, and $\epsilon_{j,t}$ is the error term at probability level τ_j and time t, such that the τ_j-th conditional quantile of $\epsilon_{j,t}$ satisfies $Q_{\epsilon_{j,t}}(\tau_j; \Omega_{t-1}) = 0$. Ω_{t-1} is the information set available at time $t-1$, with $(X_{t-1}, X_{t-2}, \ldots) \subseteq \Omega_{t-1}$. Given the multi-quantile regression model in equation 4.14, the τ_j-th conditional quantile of X_t is defined as

$$Q_{X_t}(\tau_j; \Omega_{t-1}) = \beta_0(\tau_j) + \beta_1(\tau_j)\, VaR_t(\tau_j) \quad \forall j = 1, 2, \ldots, n \tag{4.15}$$

The above equation is important since it establishes a link between the VaR forecasts and the true unknown conditional quantile from the observed losses. The backtesting procedure here is based on examining if a perfect match exists between $\text{VaR}_t(\tau_j)$ and $Q_{L_t}(\tau_j; \Omega_{t-1})$. Analogous to the hypothesis in 3.4, the test relies on regression parameters, and tests if the intercept should be 0, and intercept 1.

The backtesting procedure requires the regression coefficients to be consistently estimated. Let $\beta(\tau_j) = (\beta_0(\tau_j), \beta_1(\tau_j))'$ denote the vector of parameters for the τ_j-th quantile index, and let $\beta = (\beta(\tau_1)', \beta(\tau_2)', \ldots, \beta(\tau_n)')'$ be the stacked vector of $2n$ coefficients. Assuming the sequence $\{\tau_j, j = 1, 2, \ldots, n\}$ is ordered in the sense that $\tau_1 < \tau_2 < \ldots < \tau_n < 1$, a consistent QMLE estimator is given by

$$\widehat{\beta} = arg \min_{\beta \in \mathbb{R}^{2n}} \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{n} \rho_{\tau_j} \left(X_t - \beta_0(\tau_j) - \beta_1(\tau_j) \text{VaR}_t(\tau_j) \right) \right) \tag{4.16}$$

where $\rho_{\tau_j}(x) = x\, \psi_{\tau_j}(x)$ is the standard check function, and $\psi_{\tau_j}(x) = \tau_j - \mathbb{1}(x \leq 0)$ is the quantile step function. Under suitable regularity conditions, White et al. (2015) show that the estimator is asymptotically normally distributed:

$$\sqrt{T}\, (\widehat{\beta} - \beta) \xrightarrow{d} N\, (0, \Sigma) \tag{4.17}$$

where Σ denotes the asymptotic covariance matrix which takes the form of a Huber sandwich,

$\Sigma = A^{-1}VA^{-1}$, with:

$$V; = E(\eta_t \, \eta_t'), \tag{4.18}$$

$$\eta_t = \sum_{j=1}^{n} \nabla Q_{X_t}(\tau_j; \Omega_{t-1}) \, \psi_{\tau_j}(\epsilon_{j,t}), \tag{4.19}$$

$$A = \sum_{j=1}^{n} E[f_{j,t}(0) \, \nabla Q_{X_t}(\tau_j; \Omega_{t-1}) \, \nabla' Q_{X_t}(\tau_j; \Omega_{t-1})] \tag{4.20}$$

where $\nabla Q_{X_t}(\tau_j; \Omega_{t-1})$ denotes the $2n$ gradient vector differentiated with respect to β, and $\epsilon_{j,t} = X_t - Q_{X_t}(\tau_j; \Omega_{t-1})$,and $f_{j,t}(0)$ denotes the pdf of $\epsilon_{j,t}$ evaluated at zero.

The goal of an ES backtest is to test $\beta_0(\tau_j) = 0$ and $\beta_1(\tau_j) = 1$ for $j = 1, 2, \ldots, n$. Couperier and Leymarie (2019) propose 4 distinct null hypotheses to test various implications of these coefficient restrictions and at a much reduced number of constraints. The four backtests are denoted by J_1, J_2, I,and S and the corresponding null hypotheses H_{0,J_1}, H_{0,J_2}, $H_{0,I}$, and $H_{0,S}$ are defined as:

$$H_{0,J_1} : \sum_{j=1}^{n} (\beta_0(\tau_j) + \beta_1(\tau_j)) = n, \tag{4.21}$$

$$H_{0,J_2} : \sum_{j=1}^{n} \beta_0(\tau_j) = 0, \text{ and } \sum_{j=1}^{n} \beta_1(\tau_j) = n, \tag{4.22}$$

$$H_{0,I} : \sum_{j=1}^{n} \beta_0(\tau_j) = 0, \tag{4.23}$$

$$H_{0,S} : \sum_{j=1}^{n} \beta_1(\tau_j) = n, \tag{4.24}$$

where J_1 and J_2 indicate "joint" tests, and I, and S refer to the "intercept" backtest and "slope" backtest respectively.

The null hypotheses above have been designed to assess various implications that the regression coefficients must satisfy for the ES forecasts to be valid. The aggregation of regression coefficients over the different probability levels helps to significantly reduce the number of constraints with H_{0,J_2} characterized by two constraints, and H_{0,J_1}, $H_{0,I}$, and $H_{0,S}$ involving only a single constraint. H_{0,J_1} and H_{0,J_2} are joint backtests and look at both the intercept and slope parameters. $H_{0,I}$ and $H_{0,S}$ on the other hand focus solely on one of the two parameter components, and complement the joint backtests to identify the nature of misspecification. Rejecting $H_{0,I}$ implies that the forecasting errors are constant across time whereas rejection of $H_{0,S}$ suggests that the errors are time-varying since they change with respect to VaR predictions.

The null hypotheses can be written in the formulation of a Wald-type test

$$H_{0,W} : R_W \beta \ = \ q_W$$

where $W \in \{J_1, J_2, I, S\}$. The general expression of the test statistic can then be written as

$$W \ = \ T(R_W \widehat{\beta} \ - \ q_W)' \ (R_W \widehat{\Sigma} R_W')^{-1} \ (R_W \widehat{\beta} \ - \ q_W) \tag{4.25}$$

where T is the out of sample size, and $\widehat{\Sigma}$ denotes a consistent estimator of the asymptotic covariance matrix. Given the null hypotheses, the quantities R_W and q_W can be defined as :

- $R_{J_1} \ = \ \iota_n \otimes (1 \ \ 1) \, , \ q_{J_1} \ = \ n$
- $R_{J_2} \ = \ \iota_n \otimes \ I_2, \ \ q_{J_2} \ = \ (0 \ \ n)'$
- $R_I \ = \ \iota_n \otimes (1 \ \ 0) \, , \ q_I \ = \ 0$
- $R_S \ = \ \iota_n \otimes (0 \ \ 1) \, , \ q_S \ = \ n$

where ι_n is a n-row unit vector, and I_2 denotes an identity matrix of size 2. If the covariance matrix Σ is non-singular, using the normality condition in 4.17, the distribution of test statistics J_1, I, and S converge to a chi-squared distribution with 1 degree of freedom, and that of J_2 converges to a chi-squared distribution with 2 degrees of freedom.

4.4 Simulation study

A simulation study is presented here to compare different methods for backtesting expected shortfall. An advantage of using simulated studies instead of real data is to evaluate these backtests independent of the performance of the methods used to estimate VaR and ES. For real data, rejecting a model does not provide information on whether the rejection was correct for an incorrect model, or the rejection was incorrect for a correct ES estimating model. Using a simulated dataset would help ascertain the probability that a given model is incorrect. There are several studies available in literature where backtesting methods 1 and 2, and the asymptotic distribution have been discussed in detail. This study would focus on the third backtesting technique that uses a multi-quantile approach to backtest ES using VaR at different probability levels.

As in Chapter 3, the correct data generating process is assumed to be a zero mean, unit

unconditional variance GARCH model with Gaussian innovations:

$$R_t \;=\; \sigma_t \epsilon_t,\; t \;=\; 1, \ldots, T, \tag{4.26}$$

$$\sigma_t^2 \;=\; (1 \;-\; \alpha \;-\; \beta) \;+\; \alpha\, y_{t-1}^2 \;+\; \beta\, \sigma_{t-1}^2 \tag{4.27}$$

where $\epsilon_t \;\sim\; iid\; N(0,1)$. We consider model 1 from Section 3.3 with $\alpha \;=\; 0.05$ and $\beta \;=\; 0.9$. The true VaR and ES can then be computed as:

$$V_{t,\tau} \;=\; \sigma_t\, \Phi_\tau^{-1} \tag{4.28}$$

$$ES_{t,\tau} \;=\; \sigma_t\, \frac{\phi(\Phi^{-1}(\tau))}{1 \;-\; \tau} \tag{4.29}$$

where ϕ is the density function of the standard normal distribution, and Φ_τ^{-1} denotes the τ quantile of a standard normal distribution.

The comparative study is done to assess the performance of five methods of backtesting expected shortfall based on the approximation of ES using VaR at different levels. Simulated datasets of two different sample sizes are used $T \;=\; 500, 1000$ to examine if more data makes the testing more accurate. It is also worth examining the effect of number of VaR levels considered for approximating ES on the backtesting techniques. The study chooses $\tau = 0.975$ and the VaR levels are calculated according to equation 4.11 for $n = 1, 2, 4, 6$ levels of VaR. A special case for $n = 2$ is also considered where $\tau_1 = 0.975$, and $\tau_2 = 0.99$ as set by the Basel committee for jointly backtesting ES with VaR. It is denoted by BIS form here on.

Table 4.1 summarizes the size of 5% tests for the first set of tests. The assumption that ES can be approximated by aggregating VaR at different probability levels requires these VaR components to be tested individually or jointly. Method 1 tests these levels independently and the minimum p-value is taken as the p-value of the combined test. The null hypothesis is rejected if at least one of the n p-values is less than or equal to $1 - (1 - \alpha)^{1/n}$, where $\alpha = 0.05$ is the level of significance. The individual VaR components are backtested using the unconditional coverage test by Kupiec (1995). The empirical size of tests is higher than the nominal size for $n = 1$ but decreases as n increases until and then starts increasing again with the empirical size being closest to the nominal size for $n = 3$.

The results are different if the p-values are combined using Fisher's chi-squared statistic in 4.10. The empirical size is higher than the nominal size for $n = 1$ and increases with n. This implies that the null hypothesis is rejected more often than is expected. The p-values here are assumed to be independent and a violation of the rule may result in smaller p-values for the Chi-squared test and consequently higher empirical sizes. An alternative would be to use approximate the distribution of the statistic in 4.10 with a scaled chi-squared distribution as described in Brown (1975).

Table 4.1 Empirical size of 5% ES backtests for T=500

n	Individual Tests		Multinomial Distribution		Asymptotic Distribution
	Minimum p-value	Fisher's chi-squared	Pearson Chi-squared	Nass Approximation	
1	0.0614	0.0614	0.0426	0.0426	0.7058
2	0.0404	0.0852	0.0478	0.0404	0.3336
4	0.067	0.1194	0.0454	0.0404	0.165
6	0.0424	0.1442	0.0584	0.0448	0.104
BIS	0.0376	0.0756	0.0406	0.0356	0.2662

Component VaR at different probability levels can be tested jointly using the multinomial distribution approach in Method 2. The Pearson chi-squared test and the Nass test as described in 4.3.2 were considered to evaluate the hypotheses in 4.12. The empirical size is close to the nominal size but underestimated for different values of n. The expected number of VaR violations at higher probability levels when $T = 500$ and $n = 6$ is less than 5 and would affect the accuracy of the results for the Pearson test. The Nass test is a suggested alternative in such cases as is shown in table 4.1. While the results from the methods are comparable, the empirical size is in general higher for the Pearson's chi-squared test and the difference increases as the number of VaR components jointly tested, n, increases.

The empirical size for asymptotic distribution based test is much higher than the nominal size of 5% when compared to all other tests. The empirical size of tests however converges to the nominal size as the number of component VaRs increases. The empirical size for this test is much higher as compared to the other methods here. However, it is not clear if the inflated empirical sizes can be attributed to a higher number of VaR violations. The rejections here could have been for either component of the hypothesis but the test does not provide any information on that. The test should be further supplemented with an autocorrelation test to gain more insight.

The comparative study between the methods above has been done assuming a relatively stable market scenario as described by Model 1. The performance of these methods was found to be similar in a volatile market scenario as well (Model 2) and is thus not presented here. The multinomial quantile regression based backtesting technique described in Method 3 however performs differently in both these situations and is thus presented separately. Tables 4.2 and 4.3 report the rejection frequencies of the tests at 5% level of significance for sample sizes $T = 500$, and $T = 1000$ respectively.

The first four columns of the tables report the results of the four backtests given in 4.3.3. It is clear from the tables that using asymptotic critical values based on the chi-squared distribution induces significant size distortions. These distortions can be attributed to poor inference for

Table 4.2 Empirical size of 5% backtests using multi-quantile regression for model 1

		White approach				Koenker approach			
T	p	J_1	J_2	I	S	J_1	J_2	I	S
500	1	0.1292	0.1988	0.1366	0.14	0.067	0.1328	0.0708	0.0762
	2	0.1558	0.2402	0.162	0.1698	0.0846	0.1472	0.0874	0.0934
	4	0.1922	0.282	0.1926	0.1952	0.1054	0.1754	0.1068	0.1122
	6	0.205	0.2986	0.2048	0.2026	0.1164	0.1896	0.1164	0.1204
	BIS	0.187	0.2654	0.1894	0.1902	0.0792	0.1454	0.0848	0.0904
1000	1	0.215	0.3912	0.2956	0.2922	0.053	0.087	0.0572	0.0612
	2	0.2104	0.4578	0.3234	0.2986	0.0646	0.093	0.0674	0.0714
	4	0.2206	0.5056	0.3586	0.318	0.0716	0.0992	0.0722	0.0742
	6	0.2256	0.5234	0.3712	0.3268	0.0736	0.1054	0.0758	0.0808
	BIS	0.2168	0.481	0.3434	0.3088	.0672	0.0946	0.068	0.0722

the quantile regression parameters in the extreme tail when the sample size is not sufficiently large. It should be noted that the validity of the asymptotic inference outlined by White et al. (2015) is applicable to more central coverage levels and therefore may lead to poor inference for extreme probability levels as is the case with ES backtesting. Couperier and Leymarie (2019) also suggest the use of bootstrapped critical values for smaller sample sizes to draw meaningful conclusions.

It should also be noted that Couperier and Leymarie (2019) uses a rectangular kernel to estimate the conditional density $f_{j,t}$ of $\epsilon_{j,t}$ evaluated at zero as given by Powell et al. (1991). An estimate of matrix A in 4.18 can be written as:

$$\widehat{A} \;=\; \frac{1}{2\widehat{c}_T T} \sum_{t=1}^{T} \sum_{j=1}^{n} \mathbb{1}(|\widehat{\epsilon}_{j,t}| \leq \widehat{c}_T)\, \nabla \widehat{Q}_{X_t}(\tau_j; \Omega_{t-1})\, \nabla' \widehat{Q}_{X_t}(\tau_j; \Omega_{t-1}) \tag{4.30}$$

where $\widehat{c}_T$ is a bandwidth parameter such that $\widehat{c}_T / c_T \xrightarrow{p} 1$, and c_T is a non-stochastic positive sequence satisfying $c_T = o(1)$, and $C_T^{-1} = o(T^{1/2})$.

While the use of a kernel based approach for estimating the density function is convenient, it is best suited for central coverage levels. An alternative to this approach is to approximate the density function using the difference quotient given in 3.10. The general hypotheses

$$H_{0,W} : R_W \beta \;=\; q_W$$

Table 4.3 Empirical size of 5% backtests Using multi-quantile regression for model 2

		White approach				Koenker approach			
T	p	J_1	J_2	I	S	J_1	J_2	I	S
500	1	0.1414	0.3334	0.2578	0.284	0.0388	0.155	0.0762	0.125
	2	0.1794	0.385	0.2988	0.3114	0.0476	0.167	0.0912	0.1446
	4	0.226	0.4356	0.337	0.337	0.0642	0.2006	0.1182	0.1698
	6	0.2536	0.4628	0.3586	0.3584	0.071	0.221	0.1344	0.1804
	BIS	0.203	0.4128	0.3204	0.3256	0.045	0.1654	0.0888	0.1396
1000	1	0.1182	0.2778	0.229	0.2442	0.0384	0.1034	0.0622	0.0872
	2	0.1646	0.3066	0.245	0.2466	0.0392	0.107	0.073	0.103
	4	0.203	0.3404	0.2726	0.2578	0.0436	0.109	0.0806	0.1104
	6	0.217	0.362	0.2806	0.2758	0.0464	0.115	0.0896	0.117
	BIS	0.1832	0.3308	0.2662	0.2626	.0446	0.1092	0.076	0.1036

where $W \in \{J_1, J_2, I, S\}$ can thus be alternately tested using the test statistic given by Koenker and Bassett (1978) :

$$W \;=\; T(R_W\beta \;-\; q_W)'(R_W V_1^{-1} R_W)^{-1}\,(R_W\beta \;-\; q_W) \tag{4.31}$$

where V_1 is a $2n \times 2n$ matrix with ijth block

$$V_1(\tau_i, \tau_j) \;=\; [min(\tau_i, \tau_j) - \tau_i\tau_j]\; H(\tau_i)^{-1}\; J(\tau_i, \tau_j)\; H(\tau_j)^{-1}$$

and $J(\tau_i, \tau_j)$ and $H(\tau_i)$ are setup as in the univariate case in 3.7. The choice of h_t for the density estimate is as given in 3.15.

The last four columns in tables 4.2 and 4.3 report the empirical sizes of tests based on the test statistic given by Koenker and Bassett (1978). While the empirical size of all four tests is greater than the empirical tests, the size distortions are significantly less prominent than in the previous case. The selection of the number of risk levels n, and sample size T also affect the size of the test. The empirical size increases as n increases and decreases with an increase in T. This is in keeping with the observations by Emmer et al. (2015) on accuracy of backtests based on quantile approximations.

Tables 4.2 and 4.3 also report a comparison of the performance of these methods in two different market situations. A point of similarity in the two tables is that the empirical sizes

for Koenker's approach are much closer to the nominal size than the White approach. The size distortions are less prominent in Model 2 which may be attributed to the performance of quantile regression in the two situations. The high vol-of-vol in Model 2 aids the performance of quantile regression which reflects in these tests based on the regression parameters.

4.5 Empirical analysis

A comparative analysis is presented here assessing the performances of the tests discussed above for backtesting ES. The framework for analyzing the outcome of the backtests, and performing a comparative evaluation has three outcomes as outlined by Edberg and Kack (2017):

- Full consensus (all backtest have the same result)
- Majority consensus (a majority of the backtestshave the same results)
- No consensus (half of the backtests rejects and half of them accept)

As in Section 2.4 and 3.4, daily log returns for eight individual stocks from S&P 500 were considered. The result is presented in tables 4.4- 4.11. Two remarks should be made here. First, of the 7 methods considered here for forecasting the component VaRs, it is only possible to forecast the ES explicitly with the four EWQR methods. The formulae for these were presented in Chapter 2. Second, the ES forecasts issued from these methods is strongly correlated with the ES calculated using the quantile approximation technique irrespective of the value of n. Further, the approximation improves as n increases which is intuitive, and $n = 6$ provides a good approximation for practical purposes and is thus used for the analysis presented here.

Table 4.4 reports the p-values for the asymptotic distribution based backtest by Costanzino and Curran (2015). Rejecting the null hypothesis here could be due to the rejection of either components of the hypothesis in 4.4. The ES predictions using EWQR method are rejected for most of the stocks in the study. A possible explanation for these rejections is the unusually high number of VaR violation over all risk levels as shown in figure 4.1. On the other hand, the Adaptive CAViaR possibly rejects the ES predictions for GE and INTC due to unusually low number of violations. Also, since the test statistic depends on cumulative violations over all risk levels, the test cannot compare estimation methods with equal number of total violations as is the case of XOM for the EWDKQR leverage, and all CAViaR methods.

In Table 4.5, p-values from individual backtests for VaR at the component risk levels are comined to yield the p-value for the ES backtest using the minimum p-value approach. For $n = 6$ and $\alpha = 0.05$, the null hypothesis is rejected if any of the component p-values are less than $1 - 0.95^{1/6} = 0.0102$. This method is in majority consensus with the asymptotic distribution

approach. While this approach is easy to use, it is possible the ES backtest would be rejected only if one VaR backtest is significant at level α. It is insensitive to situations where VaR backtest for several components have small p-values but none is small enough to be significant by itself.

Tables 4.6 and 4.7 report the p-values for Pearson's chi-squared test and Nass test for testing multiple proportions. These tests are in full consensus barring a couple of instances where the p-value borders the level of significance. The instances where the ES predictions are rejected usually involve unusually high deviations from the expected number of VaR violations for the risk levels as in the case of the EWQR method. The results however should be interpreted with caution since the tests assume the violations at the levels to be independent which is not generally the case.

Tables 4.8 - 4.11 report the p-values for the multi-quantile regression analysis based backtesting for the different null hypotheses specified in 4.21. The results are based on Koenker's approach for $n = 6$ with bandwidth as given in 3.15. Since asymptotic critical values can create size distortions, p-values obtained using bootstrapped critical values are reported here. The results of the ES backtests may depend on the number of quantiles considered. Results for $n = 2$ and $n = 4$ are included in Appendix 6.1.

For the EWQR method, the null hypotheses J_1 is rejected for three stocks - JNJ, WMT, and INTC indicating that the sum of the estimated beta coefficients was different from n. Hypotheses I and S were also rejected for these stocks indicating that both the intercept and slope coefficients are different from what we expected. While the hypothesis I was rejected for several other stocks, the overall test J_1 was not rejected suggesting a possible adjustment provided by the slope parameters to the J_1 test statistic. It should also be noted that while there were a higher than expected number of VaR violations for the EWQR method, it is possible that the magnitude of violations was not as large, and hence the J_1 and S backtests were not rejected as in the case of the GE stock. The two joint tests J_1 and J_2 are in majority consensus except for the XOM stock which can be explained by the consistently high number of VaR violations at all risk levels.

The EWQR leverage method showed slight improvements over the EWQR technique with respect to backtesting VaR using the Hit, DQ, or VQR techniques for $\tau = 95\%$ as shown in earlier Chapters. The EWQR Leverage is much more conservative so that the predicted VaR is greater than the predicted VaR using the EWQR (constant) method. While this results in fewer violations, the multi-quantile regression method is able to detect departures in the predicted VaR from the conditional quantile. These departures are reflected in the estimated coefficients being far off from their expected values. Consequently, all 4 backtests reject the ES more often than the other methods considered here.

The exponentially weighted double kernel methods produce VaR estimates that are greater than the EWQR method thereby leading to fewer violations but is not as conservative as the EWQR Leverage method. The EWDKQR method is in majority consensus with the EWQR for H_{0,J_1}, and with EWQR Leverage for H_{0,J_2}. The joint backtests are rejected if either or both of $H_{0,S}$ and $H_{0,I}$ are rejected indicating if the VaR is constant or time-varying. H_{0,J_2} for EWDKQR also has a full consensus with the EWDKQR Leverage method, and a majority consensus for H_{0,J_1}. This is to be expected since the proposed modified optimization strategy uses the predicted value from the EWDKQR (constant) method as the starting value of the predicted VaR. The EWDKQR methods also perform better than the EWQR and CAViaR in terms of VaR violations with risk levels in the upper tail.

The Adaptive CAViaR technique has a high number of exceedances for several risk levels as shown in figure 4.1 and the minimum p-value results in table 4.5. Furthermore, there is almost no consensus between the minimum p-value and the asymptotic distribution approaches. The former approach is based on VaR violations at individual risk levels and the latter on cumulative violations. This indicates that while there are a higher than expected number of VaR violations at one or more risk levels, they are probably offset by the cumulative violations being in check. This is further corroborated by the joint tests in tables 4.6 and 4.7 that are in majority consensus with the asymptotic distribution approach. The results of these tests can be complemented with the joint tests based on the multi-quantile approach. The joint test J_2 is more conservative than J_1 and is more in consensus with the minimum p-value approach.

The symmetric absolute value (SAV) and asymmetric slope (AS) CAViaR approaches are in majority consensus across all methods. These methods are associated with fewer than expected VaR violations for almost all component risk levels indicating that it is possible that the VaR model is conservative. Table 2.2 showed that these tests were conservative for $\tau = 0.95$ and it seems to be the case for extreme quantiles as well.

A comparative analysis of the multi-quantile regression based backtests for $n = 2, 4,$ and 6 show that the p-values decrease with an increase in n. This indicates that the rejections are more severe when the number of risk level increases. This is in keeping with the observations in Emmer et al. (2015) where they pointed out that the power of the backtests based on a multi-quantile approximation of ES increases with the sample size T, and decreases with the number of risk levels, n.

4.6 Summary

While ES is a better risk measure than VaR in many aspect it is significantly harder to backtest. The methods discussed in this Chapter have a high degree of complexity and subjectivity arising

due to dealing with extreme quantiles. While we would like to find a method that is both simple and reliable, the decision on which backtest to use would most importantly depend on what we want to achieve with the backtest.

No method performs well under all circumstances. Method 1 or the asymptotically normal Z-test does not require sophisticated estimation methods but has comparably low abilities to reject and is not recommended unless there is no other option. Method 2 where VaR backtests are implemented independently is the simplest of the methods discussed but it is sensitive to small errors in the underlying VaR models. Further, the assumption in methods 2 and 3 to treat VaR at different risk levels as independent is unrealistic especially if n is large. Method 3 with its simplicity has the potential to be performed as a regular routine similar to the binomial distribution based Hit tests for VaR. It is also possible to construct a traffic-light system analogous to the system for VaR backtest recommended by the Basel Committee on Banking Supervision. However, the performance of the test would deteriorate as the number of risk levels increase.

Method 4 proposes a Wald-type inference for a multi-quantile regression based backtest of ES. Like the other methods, this approach is consistent with the regulatory guidance to verify the validity of the underlying ES model at levels $\tau = 0.975$ $\tau = 0.99$. The method however suffers from size distortions for smaller sample sizes, and would benefit from using a simulation or bootstrap based approach to correct these small sample biases. The number of risk levels considered is critical to the performance of all the methods discussed here. It is imminent to choose a sufficient number of quantiles to assess ES. The use of one or two risk levels in not advisable as they may not be able to identify improper risk. Using four to six risk levels on the contrary deliver more sound decisions as shown by the simulation study. These suggest an update to the regulatory guidelines in favor of evaluation of more than two quantiles to efficiently backtest ES.

Table 4.4 P-value for asymptotic distribution based normal backtest

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0001*	< 0.0001*	0.0188*	< 0.0001*	< 0.0001*	< 0.0001*	0.3868	0.0004*
EWQR Leverage	0.3432	0.0289*	0.3868	0.1492	0.0026*	0.3868	0.1738	0.0637
EWDKQR	0.0356*	0.4336	0.1738	0.5921	0.3868	0.0764	0.0356*	0.9671
EWDKQR Leverage	0.0150*	0.5921	0.1080	0.9671	0.3029	0.0094*	0.0527	0.3432
Adaptive, CAViaR	0.0026*	0.5921	0.0119*	0.5921	0.9671	0.2320	0.0058*	0.7730
SAV, CAViaR	0.1492	0.5921	0.2014	0.3029	0.1080	0.0094*	0.0234*	0.0119*
AS, CAViaR	0.0637	0.5921	0.0234	0.2658	0.0637	0.0094*	0.0356*	0.0074*

* denotes significance at $\alpha = 5\%$

Table 4.5 Testing VaR independently at all risk levels using Tippet's minimum p-value method

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	< 0.0001*	< 0.0001*	0.0096*	< 0.0001*	< 0.0001*	< 0.0001 *	0.1735	0.0003*
EWQR Leverage	0.0453	0.0745	0.0205	0.0300	0.0051*	0.0563	0.0386	0.0772
EWDKQR	0.0010*	0.2127	0.0147	0.5994	0.1390	0.0046*	0.0010*	0.3820
EWDKQR Leverage	0.0010*	0.2127	0.0058*	0.0716	0.0386	< 0.0001*	0.0046*	0.0565
Adaptive, CAViaR	< 0.0001*	0.0046*	0.001*	0.0453	0.0453	0.0046*	< 0.0001*	0.5925
SAV, CAViaR	0.0386	0.0453	0.0386	0.0147	0.0205	0.0001*	0.0005*	0.0005*
AS, CAViaR	0.0110	0.1272	0.0002*	0.0147	0.0046*	< 0.0001*	0.0005*	0.0002*

* denotes significance at $\alpha* = 1 - (1-\alpha)^{1/n}$ for $\alpha = 5\%$

Table 4.6 P-value for jointly testing component VaRs: Pearson's chi-squared test

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	< 0.0001	< 0.0001	0.0021	< 0.0001	< 0.0001	< 0.0001	0.005	0.0032
EWQR Leverage	0.2011	0.2843	0.4621	0.2760	0.1170	0.6575	0.5201	0.2843
EWDKQR	0.2700	0.0620	0.2370	0.4931	0.0051	0.4275	0.1403	0.8790
EWDKQR Leverage	0.2650	0.0446	0.3103	0.0062	0.4204	0.0803	0.3480	0.1958
Adaptive, CAViaR	0.0803	0.3117	0.2249	0.2843	0.3594	0.0498	0.1413	0.1667
SAV, CAViaR	0.4592	0.6885	0.6818	0.3194	0.4831	0.1634	0.2782	0.2731
AS, CAViaR	0.3480	0.2102	0.2650	0.5978	0.2985	0.0803	0.4026	0.1413

* denotes significance at $\alpha = 5\%$

Table 4.7 P-value for jointly testing component VaRs: Nass test

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	< 0.0001	< 0.0001	0.0045	< 0.0001	< 0.0001	< 0.0001	0.0093	0.0063
EWQR Leverage	0.2134	0.2899	0.4496	0.2824	0.1332	0.6270	0.5017	0.2899
EWDKQR	0.2769	0.0771	0.2467	0.4774	0.0094	0.4186	0.1559	0.8441
EWDKQR Leverage	0.2724	0.0583	0.3134	0.0111	0.4123	0.0964	0.3474	0.2085
Adaptive, CAViaR	0.0964	0.3146	0.2355	0.2899	0.3576	0.0641	0.1569	0.1811
SAV, CAViaR	0.4470	0.6558	0.6496	0.3217	0.4684	0.1781	0.2843	0.2797
AS, CAViaR	0.3474	0.2220	0.2724	0.5721	0.3028	0.0964	0.3964	0.1569

* denotes significance at $\alpha = 5\%$

Table 4.8 P-value for multi-quantile regression backtest (Koenker's approach): J_1

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0836	0.1156	0.11	0.0062	0.163	0.0138	0.0182	0.1014
EWQR Leverage	0.0158	0.0424	0.0328	0.0610	0.2228	< 0.0001	0.0038	0.004
EWDKQR	0.0228	0.5816	0.1382	0.033	0.1554	< 0.0001	0.0058	0.0824
EWDKQR Leverage	0.1006	0.57	0.0406	0.0966	0.492	0.0184	0.0476	0.0888
Adaptive, CAViaR	0.0596	0.0056	0.3336	0.1068	0.414	0.005	0.0922	0.039
SAV, CAViaR	0.529	0.1056	0.6426	0.452	0.7746	0.0566	0.2394	0.8752
AS, CAViaR	0.3232	0.0698	0.2842	0.7114	0.3068	0.0074	0.0994	0.8758

* denotes significance at $\alpha = 5\%$

Table 4.9 P-value for multi-quantile regression backtest (Koenker's approach): J_2

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.1104	0.0482	0.0522	0.0276	0.1472	0.022	0.056	0.1496
EWQR Leverage	0.0336	0.1048	0.0142	0.1142	0.4672	0.0008	< 0.0001	0.0286
EWDKQR	0.0198	0.8186	0.022	0.0586	0.3532	0.0008	0.0018	0.157
EWDKQR Leverage	0.0384	0.3762	0.0314	0.2074	0.5456	0.0094	0.009	0.1128
Adaptive, CAViaR	0.0008	0.0208	0.0144	0.0646	0.2956	0.0082	0.0038	0.0748
SAV, CAViaR	0.0592	0.1566	0.1264	0.0332	0.1614	0.0064	0.0638	0.1406
AS, CAViaR	0.1104	0.1564	0.0408	0.3342	0.211	0.0008	0.0176	0.0328

* denotes significance at $\alpha = 5\%$

Table 4.10 P-value for multi-quantile regression backtest (Koenker's approach): I

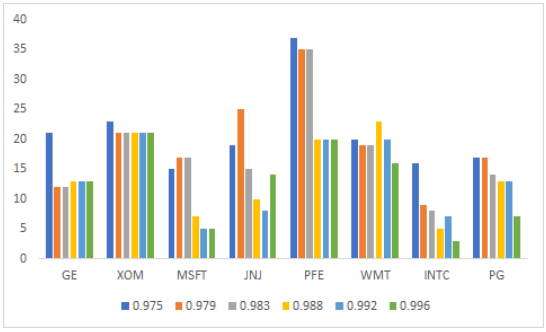
	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.044	0.035	0.0358	0.0028	0.0628	0.005	0.0316	0.071
EWQR Leverage	0.0726	0.0338	0.1494	0.0826	0.2258	0.0054	0.0380	0.0028
EWDKQR	0.171	0.559	0.3104	0.0378	0.1768	0.0066	0.0952	0.0712
EWDKQR Leverage	0.4172	0.4816	0.131	0.0978	0.6538	0.0858	0.2038	0.1214
Adaptive, CAViaR	0.9906	0.0010	0.7152	0.6370	0.5044	0.0624	0.5102	0.0470
SAV, CAViaR	0.6386	0.1824	0.834	0.3074	0.8782	0.3188	0.5776	0.6764
AS, CAViaR	0.692	0.0908	0.6426	0.5324	0.4412	0.0562	0.2244	0.4888

* denotes significance at $\alpha = 5\%$

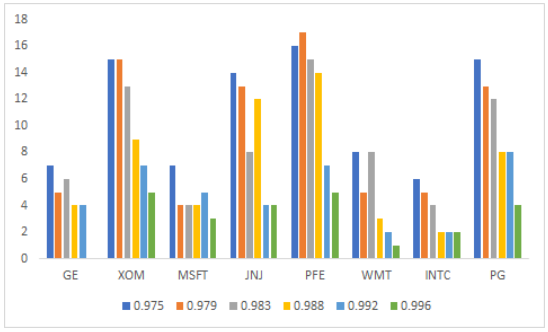
Table 4.11 P-value for multi-quantile regression backtest (Koenker's approach): S

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0816	0.113	0.1074	0.0056	0.1584	0.0132	0.0186	0.1006
EWQR Leverage	0.0162	0.0416	0.0348	0.0610	0.2220	<0.0001	0.0040	0.0040
EWDKQR	0.0236	0.5822	0.1452	0.0332	0.1556	<0.0001	0.0074	0.082
EWDKQR Leverage	0.106	0.5682	0.043	0.0968	0.498	0.0196	0.0522	0.089
Adaptive, CAViaR	0.0640	0.0052	0.3474	0.1142	0.4176	0.0054	0.1160	0.0394
SAV, CAViaR	0.557	0.1062	0.6496	0.4486	0.781	0.0598	0.2514	0.8928
AS, CAViaR	0.3322	0.07	0.294	0.7074	0.3114	0.008	0.104	0.857

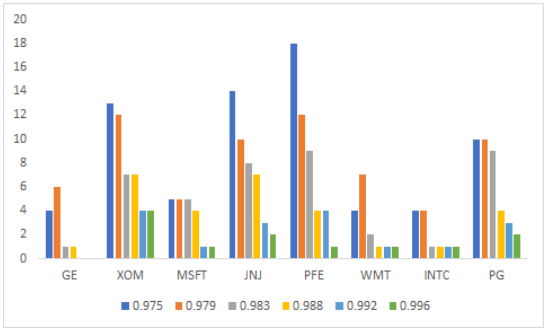
* denotes significance at $\alpha = 5\%$



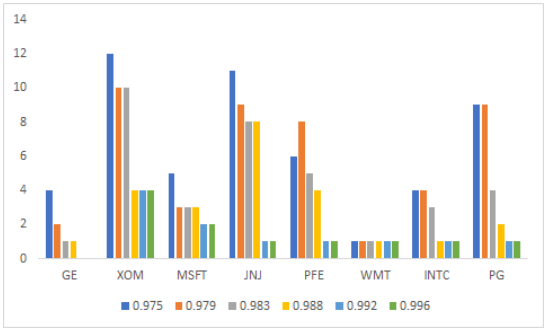
EWQR



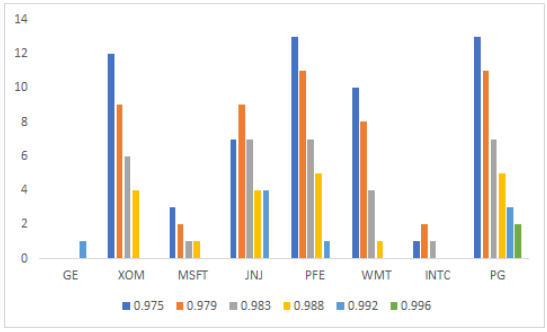
EWQR Leverage



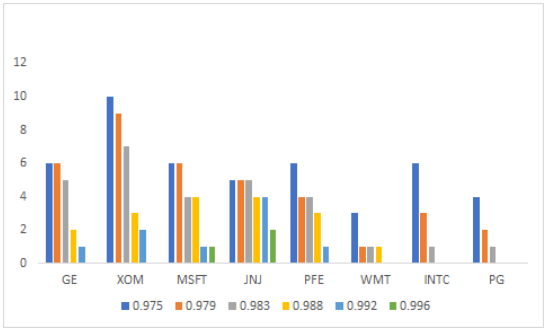
EWDKQR



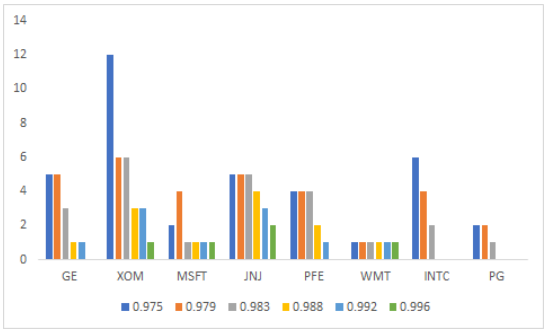
EWDKQR Leverage



Adaptive, CAViaR



SAV, CAViaR



AS, CAViaR

Figure 4.1 VaR violations at $n = 6$ probability levels for $\tau = 0.975$

CHAPTER

5

DISCUSSION AND FUTURE WORK

5.1 Discussion

Chapters 2-3 provided a comparative analysis between the EWQR and CAViaR methods with respect to three backtesting techniques - Hit, DQ, and VQR tests. The EWQR methods fared better than the CAViaR methods with the EWDKQR methods performing the best across the three tests. The Hit and DQ tests were not as conservative indicating that they may not be as powerful for finite samples. The proposed modification to the EWDKQR with leverage effect helped solve the convergence issues with the optimization problem.

Chapter 3 presented a simulation study to compare the performance of the three backtests across different sample sizes, and in different market conditions. The empirical size of the VQR test converges faster to the nominal size as the length of observations, T, increases. The power analysis showed interesting results with respect to the two models considered. While the results of the backtests for the three methods were similar in a stable market scenario, the DQ test had more power than the VQR test with same information variables in a high vol-of-vol market.

Chapter 4 presents a comparative analysis between several backtesting techniques for expected shortfall. Of the methods compared, the multi-quantile regression approach provides the most comprehensive information about the evaluation of ES forecasts. Estimating the conditional density using the sparsity based approach seemed to perform better than the kernel density method. Even though the underlying asymptotic assumptions do not hold for extreme quantiles,

modifying the bandwidth for the $\tau = 95\%$, and using the two-step approach for estimating the conditional density estimate for $\tau = 99\%$ yield good results within the current framework. Since the validity of these forecasts are tested through the validity of the component VaRs at several risk levels, the modified EWDKQR method performs the amongst the methods considered.

5.2 Future work

The results and findings from the previous chapters present many potential directions and opportunities for future work. We briefly summarize these avenues by chapter.

Quantile regression based estimation and backtesting methods for VaR and ES measures are an important contribution to the non-parametric risk analysis literature. The approaches discussed here are relatively new, and need further empirical evaluation of the methods using perhaps different data, and also a different set of benchmark estimation and backtesting methods from the literature. The thesis in the course of exploring these methods identifies the associated challenges, and suggests modifications and a working guideline to implement them. In terms of future research, it would be interesting to see a) studies addressing the inherent shortcomings of the theoretical framework, and/or b) detailed empirical evaluations to ensure objectivity in their implementation.

The exponentially weighted quantile regression VaR measure discussed in Chapter 2 assumes a linear quantile model. It may be practical in certain situations to consider a non-linear model to better capture the underlying structure. While EWQR may be used to estimate such models, the minimization problem could not be solved using linear programming. The non-linear approach can be then extended to estimate multi-period VaR or the long term VaR. Another potential area is to research the double kernel quantile regression. While the modified EWDKQR method with leverage performed well, it would be interesting to implement the original approach by Taylor (2007) of using a dynamic slope parameter with improvements to the optimization procedure.

Chapter 3 focuses on using the quantile regression framework to backtest VaR. The asymptotics associated with the quantile regression parameter are appropriate for central interval of quantiles and do not apply sufficiently far in the tails. Chernozhukov (2005) obtained the large sample properties of extremal (extreme and intermediate order) quantile regression estimators for the linear quantile regression model in both the homoscedastic and heteroscedastic cases. Given the recommendations of the Basel Committee on Banking Supervision to use a 99% VaR for regulatory purposes, it would be interesting to see how the VQR backtest can be extended based on the extremal quantile regression method. Another potential area of research for estimation and evaluation of VaR forecasts is with time aggregation or the multi-day VaR

setup. A natural extension of the VQR approach can be done using the Quantile Autoregressive (QAR) technique by Koenker and Zhao (1996). It may be possible to use the QAR setup to recursively generate multi-period density forecasts.

An added advantage of using a regression framework to base a backtest like DQ and VQR is the flexibility to add regressors. Considering additional variables in the model may help to increase the power of the VQR test in other directions. This is easier to implement in the case of a DQ test with the most common addition being VaR violation indicators at lagged time points like in Chapters 2 and 3. The VQR test on the other hand would be extended to a multivariate quantile regression framework. The extension may not be straightforward but can be built on the existing proposals in the literature that focus towards multidimensional generalization of univariate quantiles.

Chapter 4 explores different methods for backtesting ES with special reference to the multi-quantile regression setup. The most common method used for backtesting ES is the ordinary bootstrap method. However, the financial data series do not meet its assumption of i.i.d. observations and are therefore not as reliable. It is recommended to use a block bootstrap instead that would help capture the autocorrelation in the data series. While a lot of work is available on block bootstrapping for smooth functionals of weakly dependent data, not many theoretical results are available for non-smooth functions. Sun and Cheng (2018) suggested the use of a moving block bootstrap (MBB) approximation for testing expected shortfall. Further research is needed on decisions regarding the number of blocks considered, and the optimal block length.

The approach of Emmer et al. (2015) to approximate the ES using VaR at different risk levels in the tail is interesting since it does not require us to estimate ES forecast using a particular model, and still provides a way to backtest it. The joint VaR tests discussed in this thesis assumed independence for VaR series at different risk levels. The assumption however does not hold for stock data series especially for quantiles that are close to each other. This would be an issue as the number of risk levels ,n, increase. It is expected that the power of the joint test for VaR violations at these risk levels would decrease with an increase in the number of levels, n, and decrease with the sample size, T. The choice of the number of risk levels is therefore subjective and should be determined on a case by case basis with respect to sample size. It would be interesting to see an analytical approach on to decide the number of risk levels that would suffice in the linear approximation of ES by different VaRs.

The choice of n also plays a crucial role in the multi-quantile regression system. As n increases, we have an increasing number of quantiles that are far out in the tail. As with the VQR approach, the implementation of these methods require us to estimate the conditional density using the sparsity based difference quotient method. As the quantiles get closer, the chances of quantile crossing increase resulting in a larger proportion of points with conditional

density zero as given in 3.11. Using the proposed 2-step estimation method for the conditional density may offer some improvements. An alternative to this would be to use a constrained quantile estimation process to avoid crossing of quantile curves. Also, since the asymptotics discusses here are not sufficient far out in the tail, it would also be interesting to see if extremal quantile regression can be extended to multi-quantile regression setup.

BIBLIOGRAPHY

C. Acerbi and B. Szekely. Back-testing Expected Shortfall. *Risk*, 43(12):76–81, 2014.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228, 1999.

J. Berkowitz and J. O'Brien. How accurate are value-at-risk models at commercial banks? *The Journal of Finance*, 57(3):1093–1111, 2002.

J. Berkowitz, P. Christoffersen, and D. Pelletier. Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science*, 57(12):2213–2227, 2011.

E. Bofinger. Non-parametric Estimation of Density for Regularly Varying Distributions. *Australian Journal of Statistics*, (17):192–195, 1975.

M. Brown. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics*, 1975.

V. Chernozhukov. Extremal Quantile Regression. *Annals of Statistics*, 33(2):806–839, 04 2005.

P. Christoffersen. Evaluating Interval Forecasts. *International Economic Review*, 39(4):841 – 862, 1998.

P. Christoffersen. *Backtesting*. John Wiley and Sons, 2008.

P. Christoffersen and D. Pelletier. Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics*, 2(1):84–108, 2003.

P. Christoffersen, J. Hahn, and A. Inoue. Testing and Comparing Value-at-Risk Measures. *Journal of Empirical Finance*, 8(3):325–342, 2001.

R. Cont, R. Deguest, and G. Scandolo. Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10(6):593–606, 2010.

N. Costanzino and M. Curran. Backtesting General Spectral Risk Measures with Application to Expected Shortfall. *The Journal of Risk Model Validation*, 9(3):21–31, 2015.

N. Costanzino and M. Curran. A Simple Traffic Light Approach to Backtesting Expected Shortfall. *Risks*, 6(1):2, 2018.

O. Couperier and J. Leymarie. Backtesting Expected Shortfall via Multi-Quantile Regression. *halshs-01909375v2*, 2019.

M. H. A. Davis. Verification of Internal Risk Measure Estimates. *Statistics & Risk Modeling*, 33(3-4):67–93, 2016.

Z. Du and J. C. Escanciano. Backtesting Expected Shortfall:Accounting for Tail Risk. *Journal of Finance and Economics*, 43(12):74–99, 2017.

P. Edberg and B. Kack. Non-parametric Baktesting of Expected Shortfall. *KTH Royal Institute of Technology School of Engineering Sciences*, 2017.

S. Emmer, M. Kratz, and D. Tasche. What Is the Best Risk Measure in Practice? A Comparison of Standard Measures. *Journal of Risk*, 18(2):31–60, 2015.

R. F. Engle and S. Manganelli. CAViaR . *Journal of Business & Economic Statistics*, 22(4): 367–381, 2004.

W. Gaglianone, L. Lima, O. Linton, and D. Smith. Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business & Economic Statistics*, (29):150–160, 2008.

T. Gneiting. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

P. Hall and S. Sheather. On the Distribution of a Studentized Quantile. *Journal of the Royal Statistical Society: Series B (Methodological)*, (50):381–391, 1988.

D. Hendricks. Evaluation of value-at-risk models using historical data. *Economic Policy Review*, (4):39–69, 1996.

W. Hendricks and R. Koenker. Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity. *Journal of the American Statistical Association*, 87(417):58–68, 1991.

G. A. Holton. Defining Risk. *Financial Analysts Journal*, 6(60):19–25, 2004.

P. J. Huber. The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 221–233, 1967.

M. Jones and P. Hall. Mean Squared Error Properties of Kernel Estimates or Regression Quantiles. *Statistics & Probability Letters*, 10(4):283–289, 1990.

P. Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 2007.

R. W. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. W. Koenker and G. Bassett. Regression Quantiles. *Econometrica*, 46(1):33–50, 1978.

R. W. Koenker and Q. Zhao. *Conditional Quantile Estimation and Inference for ARCH Models*, volume 12. Econometric Theory, 1996.

M. Kratz, Y. Lok, and A. McNeil. Multinomial VAR Backtests: A Simple Implicit Approach to Backtesting Expected Shortfall. *Essec Working Paper 1617*, 2016.

P. H. Kupiec. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84, 1995.

N. S. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138. ACM, 2008.

A. J. McNeil and R. Frey. Estimation of Tail-related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach . *Journal of Empirical Finance*, 7(3-4):271–300, 2000.

F. R. McNeil, Alexander J. and P. Embrechts. *Quantitative Risk Management*. Princeton University Process, 2005.

K. Osband and S. Reichelstein. Information-eliciting Compensation Schemes. *Journal of Public Economics*, 27(1):107–115, 1985.

M. P. Richardson, J. Boudoukh, and R. F. Whitelaw. The best of both worlds: A hybrid approach to calculating value at risk. *RISK*, 11, 11 1997.

J. L. Powell, W. A. Barnett, and G. E. Tauchen. Estimation of Monotonic Regression Models Under Quantile Restrictions. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 1991.

RiskMetrics. Technical document.

M. Siddiqui. Distribution of Quantiles in Samples from a Bivariate Population. *Journal of Research of the National Bureau of Standards*, 64:145–150, 1960.

S. Sun and F. Cheng. Bootstrapping the Expected Shortfall. *Theoretical Economics Letters*, 8 (1):685–698, 2018.

J. Taylor. Using Exponentially Weighted Quantile Regression to Estimate Value at Risk and Expected Shortfall. *Journal of Financial Econometrics*, 6(3):382–406, 2007.

H. White, T. Kim, and S. Manganelli. VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics*, 187(1):169 – 188, 2015. ISSN 0304-4076.

K. Yu and M. C. Jones. Local Linear Quantile Regression. *Journal of the American Statistical Association*, 93(441):228–237, 1998.

CHAPTER

6

APPENDIX

6.1 P-value for ES backtests for $n = 2$

Table 6.1 P-value for asymptotic distribution based normal backtest for n=2

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0001*	<0.0001*	0.0188*	<0.0001*	<0.0001*	<0.0001*	0.3868	0.0004*
EWQR Leverage	0.3432	0.0289*	0.3868	0.1492	0.0026*	0.3868	0.1738	0.0637
EWDKQR	0.0356*	0.4336	0.1738	0.5921	0.3868	0.0764	0.0356*	0.9671
EWDKQR Leverage	0.0150*	0.5921	0.1080	0.9671	0.3029	0.0094*	0.0527	0.3432
Adaptive, CAViaR	0.0026*	0.5921	0.0119*	0.5921	0.9671	0.2320	0.0058*	0.7730
SAV, CAViaR	0.1492	0.5921	0.2014	0.3029	0.1080	0.0094*	0.023*4	0.0119*
AS, CAViaR	0.0637	0.5921	0.0234*	0.2658	0.0637	0.0094*	0.0356*	0.0074*

* denotes significance at $\alpha = 5\%$

Table 6.2 P-value for testing VaR independently at all risk levels for n=2

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0096*	<0.0001*	<0.0001*	<0.0001*	0.1735	0.0003*
EWQR Leverage	0.0453	0.0745	0.0205*	0.0300	0.0051*	0.0563	0.0386	0.0772
EWDKQR	0.0010*	0.2127	0.0147*	0.5994	0.1390	0.0046*	0.0010*	0.3820
EWDKQR Leverage	0.0010*	0.2127	0.0058*	0.0716	0.0386	<0.0001*	0.0046*	0.0565
Adaptive, CAViaR	<0.0001*	0.0046*	0.001*	0.0453	0.0453	0.0046*	<0.0001*	0.5925
SAV, CAViaR	0.0386	0.0453	0.0386	0.0147*	0.0205*	0.0001*	0.0005*	0.0005*
AS, CAViaR	0.0110	0.1272	0.0002*	0.0147	0.0046*	<0.0001*	0.0005*	0.0002*

* denotes significance at $\alpha* = 1-(1-\alpha)^{1/2}$ for $\alpha = 5\%$

Table 6.3 P-value for jointly testing component VaRs: Pearson's chi-squared test for n=2

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0021*	<0.0001*	<0.0001*	<0.0001*	0.005*	0.0032*
EWQR Leverage	0.2011	0.2843	0.4621	0.2760	0.1170	0.6575	0.5201	0.2843
EWDKQR	0.2700	0.0620	0.2370	0.4931	0.0051*	0.4275	0.1403	0.8790
EWDKQR Leverage	0.2650	0.0446*	0.3103	0.0062*	0.4204	0.0803	0.3480	0.1958
Adaptive, CAViaR	0.0803	0.3117	0.2249	0.2843	0.3594	0.0498*	0.1413	0.1667
SAV, CAViaR	0.4592	0.6885	0.6818	0.3194	0.4831	0.1634	0.2782	0.2731
AS, CAViaR	0.3480	0.2102	0.2650	0.5978	0.2985	0.0803	0.4026	0.1413

* denotes significance at $\alpha = 5\%$

Table 6.4 P-value for jointly testing component VaRs: Nass test

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0045*	<0.0001*	<0.0001*	<0.0001*	0.0093*	0.0063*
EWQR Leverage	0.2134	0.2899	0.4496	0.2824	0.1332	0.6270	0.5017	0.2899
EWDKQR	0.2769	0.0771	0.2467	0.4774	0.0094*	0.4186	0.1559	0.8441
EWDKQR Leverage	0.2724	0.0583	0.3134	0.0111*	0.4123	0.0964	0.3474	0.2085
Adaptive, CAViaR	0.0964	0.3146	0.2355	0.2899	0.3576	0.0641	0.1569	0.1811
SAV, CAViaR	0.4470	0.6558	0.6496	0.3217	0.4684	0.1781	0.2843	0.2797
AS, CAViaR	0.3474	0.2220	0.2724	0.5721	0.3028	0.0964	0.3964	0.1569

* denotes significance at $\alpha = 5\%$

Table 6.5 P-value for ES backtests for $n = 2$ (Koenker's approach): J_1

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.27	0.6044	0.3554	0.039*	0.4134	0.0696	0.3558	0.3734
EWQR Leverage	0.1034	0.3518	0.363	0.4408	0.6028	0.0142*	0.0472*	0.041*
EWDKQR	0.1322	0.6366	0.3666	0.148	0.5306	0.042*	0.2404	0.4844
EWDKQR Leverage	0.4318	0.8476	0.5164	0.1256	0.571	0.1904	0.3292	0.3024
Adaptive, CAViaR	0.3192	0.0544	0.4832	0.3522	0.6566	0.0354*	0.6192	0.0346*
SAV, CAViaR	0.459	0.5124	0.9678	0.8514	0.6878	0.3248	0.7778	0.6362
AS, CAViaR	0.8018	0.4452	0.8394	0.8622	0.5556	0.0848	0.441	0.6816

* denotes significance at $\alpha = 5\%$

Table 6.6 P-value for multi-quantile regression backtest for $n = 2$ (Koenker's approach): J_2

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.1812	0.4726	0.629	0.0866	0.5386	0.1482	0.4824	0.6594
EWQR Leverage	0.1994	0.621	0.5368	0.6698	0.8184	0.0224*	<0.0001*	0.0224*
EWDKQR	0.1174	0.8822	0.2134	0.2928	0.7898	0.039*	0.1314	0.7298
EWDKQR Leverage	0.0214*	0.886	0.4456	0.198	0.6552	0.1512	0.171	0.516
Adaptive, CAViaR	0.0198*	0.0044*	0.0184*	0.4634	0.7612	0.0642	0.0536	0.1136
SAV, CAViaR	0.3392	0.6354	0.7504	0.5902	0.5462	0.1324	0.6444	0.426
AS, CAViaR	0.396	0.721	0.0772	0.8468	0.3512	0.01*	0.2734	0.2292

* denotes significance at $\alpha = 5\%$

Table 6.7 P-value for multi-quantile regression backtest for $n = 2$ (Koenker's approach): I

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.1576	0.4138	0.3878	0.0566	0.2936	0.049*	0.2894	0.3812
EWQR Leverage	0.1936	0.3678	0.4898	0.3878	0.553	0.0586	0.3002	0.0846
EWDKQR	0.3808	0.6282	0.6294	0.1694	0.5794	0.1462	0.5578	0.5018
EWDKQR Leverage	0.5944	0.9026	0.7116	0.1596	0.6726	0.301	0.7152	0.341
Adaptive, CAViaR	0.9856	0.1056	0.7538	0.3914	0.7118	0.0548	0.718	0.0306*
SAV, CAViaR	0.5924	0.6164	0.904	0.7842	0.7874	0.5444	0.8808	0.5048
AS, CAViaR	0.9614	0.4548	0.8794	0.7842	0.74	0.2758	0.5888	0.5264

* denotes significance at $\alpha = 5\%$

Table 6.8 P-value for multi-quantile regression backtest for $n = 2$ (Koenker's approach): S

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.2676	0.5994	0.3564	0.039*	0.407	0.0682	0.3534	0.3732
EWQR Leverage	0.1052	0.3522	0.3672	0.4388	0.6008	0.0144*	0.0514	0.0416*
EWDKQR	0.1354	0.636	0.3752	0.1478	0.532	0.0432*	0.252	0.4846
EWDKQR Leverage	0.434	0.8486	0.5228	0.1262	0.573	0.1924	0.3432	0.303
Adaptive, CAViaR	0.3322	0.055	0.498	0.3532	0.6594	0.0362*	0.6524	0.0346*
SAV, CAViaR	0.4618	0.5138	0.9732	0.8478	0.6908	0.3308	0.7818	0.632
AS, CAViaR	0.8066	0.4456	0.8524	0.8604	0.5638	0.0896	0.448	0.6776

* denotes significance at $\alpha = 5\%$

6.2 P-value for ES backtests for $n = 4$

Table 6.9 P-value for asymptotic distribution based normal backtest for n=4

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0188*	<0.0001*	<0.0001*	<0.0001*	0.2568	0.0013*
EWQR Leverage	0.7107	0.0065*	0.3868	0.0637	0.0356*	0.3868	0.2163	0.0260*
EWDKQR	0.0134*	0.2163	0.2658	0.3226	0.2658	0.0479*	0.0637	0.6209
EWDKQR Leverage	0.0479*	0.7107	0.2658	0.8047	0.9016	0.0094*	0.0260*	0.8047
Adaptive, CAViaR	0.0045*	0.8047	0.0188*	0.2658	0.7107	0.3868	0.0045*	0.7107
SAV, CAViaR	0.1080	0.7107	0.2163	0.2658	0.1738	0.0134*	0.0479*	0.0188*
AS, CAViaR	0.0637	0.7107	0.0188*	0.3226	0.0835	0.0094*	0.0356*	0.0094*

* denotes significance at $\alpha = 5\%$

Table 6.10 P-value for testing VaR independently at all risk levels for n=4

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0268	<0.0001*	<0.0001*	<0.0001*	0.3362	0.0128
EWQR Leverage	0.0862	0.0964	0.0418	0.03	0.2733	0.1054	0.0386	0.0165
EWDKQR	0.0004*	0.5817	0.0147	0.6733	0.1390	0.0046*	0.0046*	0.2773
EWDKQR Leverage	0.0046*	0.3820	0.0147	0.1784	0.1784	<0.0001*	0.0004*	0.0565
Adaptive, CAViaR	<0.0001*	0.1784	0.0004*	0.0142	0.1784	0.0110*	<0.0001*	0.5375
SAV, CAViaR	0.0142	0.1727	0.0386	0.0147	0.0386	0.0004*	0.0005*	0.0005*
AS, CAViaR	0.0110*	0.1727	0.0002*	0.0147	0.0046*	<0.0001*	0.0005*	0.0002*

* denotes significance at $\alpha* = 1 - (1 - \alpha)^{1/4}$ for $\alpha = 5\%$

Table 6.11 P-value for jointly testing component VaRs: Pearson's chi-squared test for n=4

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0981	0.0006*	<0.0001*	<0.0001*	0.1309	0.0378*
EWQR Leverage	0.3606	0.2712	0.4009	0.7988	0.7355	0.6715	0.4009	0.0223*
EWDKQR	0.0940	0.9465	0.1492	0.6782	0.0367*	0.1675	0.1094	0.6531
EWDKQR Leverage	0.2464	0.4425	0.4438	0.1901	0.7225	0.0254*	0.0923	0.0218*
Adaptive, CAViaR	0.0394*	0.3095	0.0795	0.2385	0.7225	0.1769	0.0259*	0.4419
SAV, CAViaR	0.3550	0.7814	0.4757	0.2338	0.4236	0.0795	0.1876	0.1295
AS, CAViaR	0.2464	0.3152	0.0530	0.2623	0.1675	0.0254*	0.1326	0.0427*

* denotes significance at $\alpha = 5\%$

Table 6.12 P-value for jointly testing component VaRs for n=4: Nass test

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	<0.0001*	<0.0001*	0.0045*	<0.0001*	<0.0001*	<0.0001*	0.0093*	0.0063*
EWQR Leverage	0.2134	0.2899	0.4496	0.2824	0.1332	0.6270	0.5017	0.2899
EWDKQR	0.2769	0.0771	0.2467	0.4774	0.0094*	0.4186	0.1559	0.8441
EWDKQR Leverage	0.2724	0.0583	0.3134	0.0111*	0.4123	0.0964	0.3474	0.2085
Adaptive, CAViaR	0.0964	0.3146	0.2355	0.2899	0.3576	0.0641	0.1569	0.1811
SAV, CAViaR	0.4470	0.6558	0.6496	0.3217	0.4684	0.1781	0.2843	0.2797
AS, CAViaR	0.3474	0.2220	0.2724	0.5721	0.3028	0.0964	0.3964	0.1569

* denotes significance at $\alpha = 5\%$

Table 6.13 P-value for multi-quantile regression backtest for $n = 4$ (Koenker's approach): J_1

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0996	0.3072	0.0864	0.0292*	0.1138	0.0244*	0.1728	0.0398*
EWQR Leverage	0.0304*	0.0688	0.2084	0.3018	0.4410	0.0036*	0.0126*	0.0416*
EWDKQR	0.2136	0.5990	0.1106	0.0424*	0.28	0.0014*	0.0260*	0.2586
EWDKQR Leverage	0.1648	0.6914	0.1284	0.0634	0.4814	0.0080*	0.1110	0.2464
Adaptive, CAViaR	0.0804	0.0564	0.4306	0.2256	0.5870	0.0292*	0.5108	0.0382*
SAV, CAViaR	0.3050	0.1940	0.7824	0.6136	0.7736	0.0766	0.5376	0.9124
AS, CAViaR	0.4942	0.1624	0.5088	0.8528	0.4162	0.0382	0.1826	0.7106

* denotes significance at $\alpha = 5\%$

Table 6.14 P-value for multi-quantile regression backtest for $n = 4$ (Koenker's approach): J_2

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.1634	0.0898	0.1398	0.0734	0.2768	0.0600	0.3738	0.1252
EWQR Leverage	0.0506	0.1564	0.2992	0.5600	0.7418	0.0094	0.0030	0.0792
EWDKQR	0.0324*	0.8610	0.0196*	0.0714	0.5460	0.0004*	0.0156*	0.5218
EWDKQR Leverage	0.1028	0.8386	0.0332*	0.1262	0.6632	0.0188*	0.0156*	0.3310
Adaptive, CAViaR	0.0076*	0.0716	0.0162*	0.1840	0.4242	0.0076*	0.0158*	0.1246
SAV, CAViaR	0.1382	0.2616	0.5028	0.1350	0.2672	0.0152*	0.2990	0.1582
AS, CAViaR	0.0170*	0.3090	0.1348	0.8894	0.3242	0.0038*	0.0260*	0.2712

* denotes significance at $\alpha = 5\%$

Table 6.15 P-value for multi-quantile regression backtest for $n = 4$ (Koenker's approach): I

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0550	0.1206	0.1242	0.0288*	0.1346	0.0150*	0.1790	0.0402*
EWQR Leverage	0.1248	0.0816	0.3100	0.3286	0.4532	0.0126*	0.1100	0.0250*
EWDKQR	0.5960	0.6006	0.3030	0.0482*	0.3	0.0332	0.1932	0.2602
EWDKQR Leverage	0.3564	0.6468	0.2904	0.0736	0.6146	0.0484*	0.4374	0.2866
Adaptive, CAViaR	0.8064	0.0240*	0.7200	0.2836	0.7518	0.3862	0.4754	0.0362*
SAV, CAViaR	0.7320	0.3212	0.9246	0.4662	0.8920	0.3304	0.8248	0.8228
AS, CAViaR	0.0170*	0.2326	0.8218	0.9086	0.5928	0.1918	0.3612	0.5756

* denotes significance at $\alpha = 5\%$

Table 6.16 P-value for multi-quantile regression backtest for $n = 4$ (Koenker's approach): S

	GE	Exxon	MS	J&J	Pfizer	Wal-Mart	Intel	P&G
EWQR	0.0980	0.3016	0.0876	0.0286*	0.1134	0.0244*	0.1726	0.0400*
EWQR Leverage	0.0318*	0.0692	0.2108	0.3024	0.4410	0.0038*	0.0134*	0.0412*
EWDKQR	0.2224	0.5976	0.1162	0.0424*	0.2794	0.0014*	0.0296*	0.2594
EWDKQR Leverage	0.1668	0.6916	0.1334	0.0642	0.4846	0.0084*	0.1204	0.2472
Adaptive, CAViaR	0.0870	0.0544	0.4424	0.2284	0.5954	0.0338*	0.5770	0.0384*
SAV, CAViaR	0.3152	0.1964	0.7876	0.6074	0.7798	0.0802	0.5496	0.9242
AS, CAViaR	0.5074	0.1630	0.5244	0.8524	0.4248	0.0410*	0.1898	0.7054

* denotes significance at $\alpha = 5\%$