Statistical Analysis of the Log Returns of Financial Assets

Leo Quigley

Student ID Number: 0442372

BSc in Financial Mathematics

Supervisor: Dr. Volkert Paulsen

Second Reader: Dr. David Ramsey

April 9, 2008

Abstract

In many models of financial mathematics, such as the mean-variance model for portfolio selection and asset pricing models, the independence and identical normal distribution of the asset returns is the cornerstone assumption on which these are built. Empirical studies have shown that the returns of an asset don't actually follow a normal distribution but in fact they have fatter tails than the normal can capture. There is evidence that the asset returns not only display this so-called heavy tailed behaviour but are also possibly skewed in their distributions. Empirical research has also found that returns display alternating periods of high and low volatility contradicting the idea of independent and identical distribution.

Acknowledgments

I would like to thank my family for all of their support throughout my time in college.

Thanks to everyone who I have become friends with in college who have made this the best four years of my life.

I would also like to thank my supervisor, Dr. Volkert Paulsen, for both suggesting this topic and assisting me throughout.

Contents

Chapter 1

Introduction

1.1 Objectives

During the course of this paper we will investigate the log return data of a number of financial assets. It is the aim of this project to discover whether the log return data displays certain properties of well known parameterised distributions. This will be achieved by comparing the statistical properties and characteristics of the empirical data under study to the theoretical distributions we suspect it might come from. These properties and characteristics will be assessed both through graphical and numerical methods to give a well-formed insight into the data.Using this knowledge obtained from these procedures I will then investigate whether the data can be fitted to a known distribution using fitting methods including maximum likelihood estimation.

This paper will address two main questions:

1. Are the log returns of the financial data normally distributed?

2. Are these same log returns independent and identically distributed?

1.2 Outline of Paper

For the purposes of this study we will examine the Dow Jones Industrial Average Index (NYSE:DJI) and five publically quoted companies stocks for my study. These companies are Boeing (NYSE:BA), Citigroup (NYSE:C), General Motors (NYSE:GM), Intel (NasdaqGS:INTC) and Wal-Mart (NYSE:WMT). The data was downloaded from the historical prices page on the Yahoo finance website taking the closing prices of these stocks at three different regular time intervals, specifically monthly, weekly and daily. The data was downloaded in spreadsheet form and imported to the statistical software package R on which most of the analysis was carried out. This software is open source and is a free download at http://www.r-project.org with supplementary packages available from http://cran.r-project.org. The adjusted closing prices allowing for stock splits and dividend payments were taken as the base stock value.

In chapter 2 an overview will be given into the properties of financial data. The common assumptions regarding empirical assets price trends and the nature of returns on financial assets will be discussed.

In chapter3 we will first carry out some exploratory data analysis on the monthly stock data of the chosen companies. This will involve examining the plots of the raw data, the log returns and using tools such as the Q-Q

plots to compare the sample data to simulated data that follows the normal distribution. As further checks for normality in the data we will use statistical tests such as the Jarque-Bera test, the Shapiro-Wilk test and the Kolmogorov-Smirnov tests.

In chapter 4 we will look at the tails of the distribution, in particular the tail of losses and through the under the practice of extreme value we will apply the Peak Over Threshold method. Through the introduction of the Pickands, Balkema and de Haan theorem it will be suggested that if the returns are heavy tailed that a generalised Pareto distribution will be suitable to model the data. Once these concepts are introduced we will try to apply them to the empirical data.

Next we will investigate the affect that the independence assumption failing will have on the models. We will look at the time dependency of the returns and introduce time series analysis and the various ideas that it incorporates. The concepts of a stationary time series, autocorrelation, white noise and linear time series will be discussed. Then using methods such as the autocorrelation function plot we will examine in greater detail the dependence of asset returns.

Finally, in chapter 6 findings of the project will be discussed and conclusions drawn.

Chapter 2

Overview of Returns of Financial Assets

When examining financial assets it is most common to study the returns rather than the actual raw asset prices. The reasons for analysing the returns rather than the asset price are that they give us a scale-free assessment of the performance of the asset and that returns also have more attractive statistical properties for analysis. However, it is important to note that there are many different definitions of financial assets returns.

2.1 Properties of Stock Prices

For a brief insight into the data underlying returns we will first look at some of the plots of the raw and the log of the raw data. As an example the plots for Boeing over time are shown in Figures 2.1 and 2.2. As can be seen from

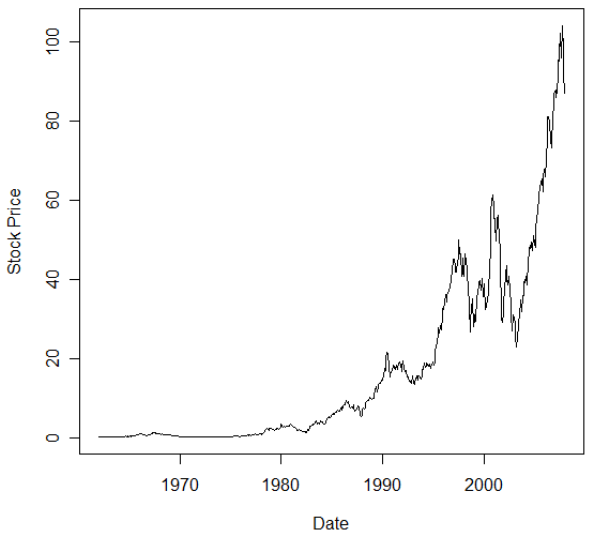

Figure 2.1: Plot of Boeing Monthly Stock Prices

Figure 2.1 the plot the stock prices display a roughly exponential growth over the full time period. This is supported by the almost linear pattern of the log of the stock prices. There are a number of key issues that should be raised. The very low price of the stock in the sixties and seventies has led to negative or only slightly positive values of the log prices. This was shortly after the company's Initial Public Offering on the stock exchange. Another obvious characteristic in the plot of the Boeing is the severe plummet in value of the stock price after the year 2000. This drop is due to a frightened market associated with the 9-11 terrorist attacks in the U.S of 2001 and because of the nature of Boeings' airline business its value was hit particularly hard by this uncertainty in the marketplace and the airline industry.

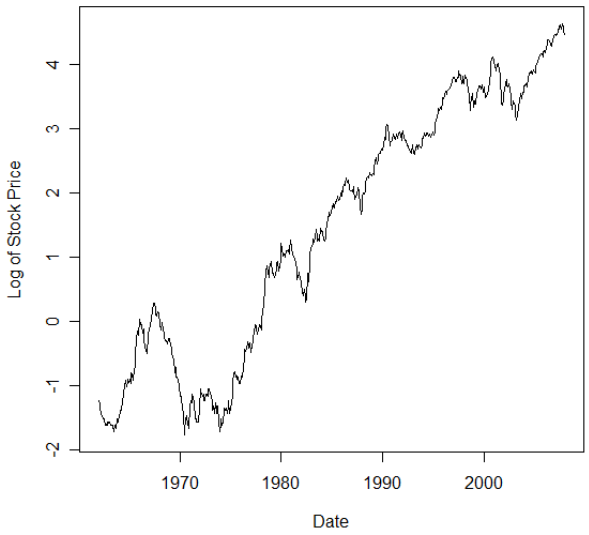

Figure 2.2: Plot of Boeing Monthly Log Prices

We can definitely come to the conclusion that the asset prices and the logs of the asset prices observed at a monthly frequency are not normally distributed. We will now concentrate solely on the returns of the assets. As these are relational to the prices immediately previous to them they are less affected by the length of the time scale involved and therefore give a better scaled reference to a stocks' performance over time.

As previously stated the log returns are often assumed i.i.d. normal. In the following sections we will rest this assumption. First we will examine the basic plots of the monthly returns to see if it can be reasonably possible to test for normality. In Figures 3.3 and 3.4 are the log returns of Citigroup and the Dow Jones index respectively. The patterns of these plots tend to

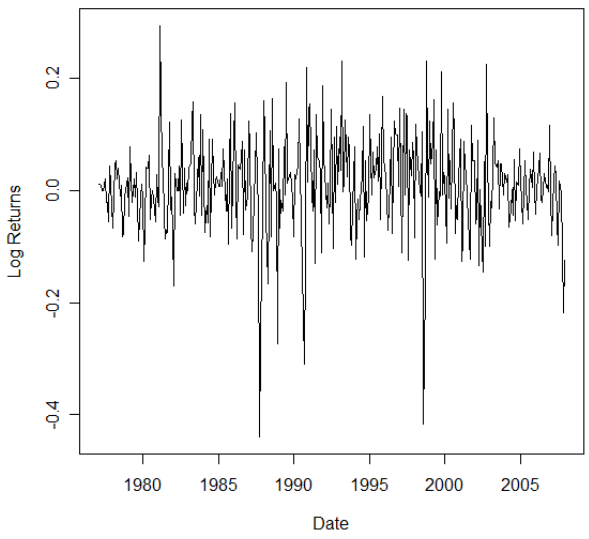

Figure 2.3: Plot of Citigroup Monthly Log Prices

fluctuate about a mean close to 0. The returns seem to be more suitable to be modelled using the random walk approach. Initially it seems it will be hard to reject the hypothesis of the returns being non-normal.

2.2 Defining a Financial Asset Return

To define the most common types of returns let S_t be the price of an asset at a time index t and assume that there are no dividends paid unless told otherwise.

A one-period simple return is the relative change in net asset price over one period from time $t-1$ to time t. The gross one-period simple return is

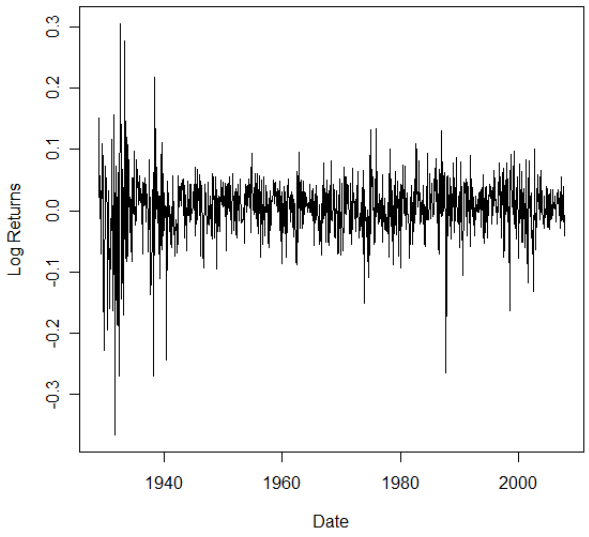

Figure 2.4: Plot of Dow Jones Monthly Log Returns

defined by

$$1 + R_t = \frac{S_t}{S_{t-1}}$$

where R_t is said to be the net one-period simple return.

A multi-period return is the return on an asset that is held for more than one period. A gross multi-period simple return over k time periods is defined as

$$1 + R_t[k] = \frac{S_t}{S_{t-k}}$$

This is also known as a compound return. $R_t[k]$ is called the k-period simple net return and is given by

$$R_t[k] = \frac{S_t - S_{t-k}}{S_{t-k}}$$

A continuously compounded return is defined as the natural logarithm of the gross simple return of the asset and is given by

$$\begin{aligned} r_t &= ln(1+R_t) \\ &= ln\left(\frac{S_t}{S_{t-1}}\right) \\ &= lnS_t - lnS_{t-1} \end{aligned}$$

This is the difference between the natural log of the assets price at time t and the natural log of its price at the previous step in time. Due to this definition r_t is also commonly called the log return of an asset.

Log returns have some more favourable properties for statistical analysis than the simple net returns R_t. The continuously compounded multi-period return is simply the sum of the continuously compounded one-period returns as shown below

$$\begin{aligned} r_t[k] &= ln(1+R_t[k]) \\ &= ln[(R_t)(1+R_{t-1})(1+R_{t-2})\ldots(1+R_{t-k+1})] \\ &= ln(1+R_t) + ln(1+R_{t-1}) + \ldots + ln(1+R_{t-k+1}) \\ &= r_t + r_{t-1} + \ldots + r_{t-k+1} \end{aligned}$$

Also the statistical properties of log returns are better behaved than simple returns.

Another variation on a return is what is termed an excess return. The excess return at time t is the difference between the assets return and that of another reference asset. Excess returns can be calculated for both simple returns and log returns as shown below:

$$\text{Simple Excess Return at time } t: \qquad Z_t = R_t - R_{0t}$$
$$\text{Log Excess Return at time } t: \qquad z_t = r_t - r_{0t}$$

where R_{0t} and r_{0t} are the simple and log returns of the reference asset. The reference assets may be anything but it is often a riskless asset such as a short-term Treasury Bill.

Also worth noting is the effect a dividend payment will have on the return. If a dividend D_t is paid between times $t-1$ and t then the return for time t needs to be adjusted to allow for the value drop after payment date. To accomplish this, the value of the dividend payment is added to the price of the asset at time t and this value is used as the real value of the asset for time t. The normal formula for working out the various returns can then be applied. For instance the log return at time t when a dividend D_t was paid between time t and $t-1$ is given by

$$r_t = ln(S_t + D_t) - ln(S_{t-1})$$

Although there are a wide variety of returns we will restrict our studies to log returns. They are the most popular form of returns to be studied when investigating financial assets. For the duration of this project log returns shall simply be referred to as returns while all other forms shall retain their full titles.

2.3 Statistical Properties of Returns

It is commonly assumed that the simple returns are log-normally distributed. A log-normal distribution has a short lower tail and a fatter upper tail. If simple returns are independent and identically distributed (i.i.d.) as log-normal then it follows that the log-returns are i.i.d. normally distributed which allows great statistical freedom. As with life everything is not so clean cut. Unfortunately there are a number of points that initially discourage acceptance of the idea of returns being independent and identically normally distributed. Firstly the lower bound of a simple return, defined as

$$R_t = \frac{S_t - S_{t-1}}{S_{t-1}} = \frac{S_t}{S_{t-1}} - 1$$

is -1. Therefore from its definition:

$$r_t = ln(1 + R_t)$$

the log return also has a lower bound whereas a normally distributed random variable has no lower bound as it can take any real value on the line. Secondly, empirical assets studied have shown kurtosis higher than the normal distribution giving the distribution of the returns heavy tails, also called extra- or excess-kurtosis.

Other important statistical properties that empirical studies have raised questions about are the skewness and kurtosis of the returns' distribution. Skewness is the normalised third central moment and it describes the symmetry of the random variable with respect to its mean. Kurtosis is the normalised fourth central moment and it describes the behaviour of the tail

of the distribution. It is independent of scale and location parameters and so can be used as a comparison coefficient between the empirical data and the normal distribution. Together skewness and kurtosis summarise the extent of asymmetry and tail thickness of the distribution.

The kurtosis of a normal distribution is 3 and a kurtosis figure of higher than this indicates that the data is fat tailed or leptokurtic. Excess kurtosis is defined as being the difference between the kurtosis of the distribution of interest and that of the normal distribution. Mathematically it can be written $K(x) - 3$ where $K(x)$ is the kurtosis of the empirical data. From empirical studies of financial data one of the properties there is strong evidence to support is this heavy tailed behaviour. Heavy tailed behaviour indicates a greater probability of extreme values being realised. As there is a larger area in these fat tails of the probability distribution curve they have the effect of reducing the cumulative probability around the centre of the curve. This affects the variance of the distribution or, as it is commonly called in financial data analysis, the volatility. Observations that follow a heavy tail distribution also destroy other classical statistical procedures such as the sample mean.

Volatility is the conditional variance of the asset return. It is the main measure used to quantify the riskiness of an asset. From empirical research it has been observed that volatility tends to increase with the square root of time as time increases and it increases in a continuous manner with it being rare to observe volatility jumps. Volatility is usually stationary meaning it varies within a fixed range and is therefore finite. There are a number of

different types of volatility such as historical volatility, which is calculated from the past returns of the asset or implied volatility, which is inversely obtained from a model that has been accepted. It is worth noting that volatility does not actually imply a direction of dispersion.

Volatility clustering is the phenomenon of spells of high amplitude that alternate with spells of low amplitude. That is to say that if at a time t the volatility is high then at the next consecutive time $t + 1$ the return will also tend to have high volatility. This characteristic contradicts the independent and identically distributed assumption that is traditionally assumed to be a stylized fact of log returns. Empirical studies of returns have shown that extreme values do not normally appear unpredictably but rather they happen after the occurrence of larger than normal movements in the return value. Periods of high volatility in returns will commonly follow a 'crash versus bubble' pattern swinging from higher than average positive values to lower than average negative values and back again. During periods of low volatility, also known as the doldrums, the returns stay much closer to the mean value with little deviation.

Chapter 3

Random Walk Approach and Normality of Returns

3.1 Random Walk Hypothesis

A random walk is defined as a process where the value of the variable of interest at a certain time depends only on its value at some previous time and a random variable which determines the step size and direction. The random walk hypothesis is a popular way to describe the evolution of stock prices, in particular the log prices. It assumes the stock price follows a stochastic process in which it moves up or down from the previous value at random. Although values in a stochastic process may be independent random variables, in most commonly considered applications such as the evolution of asset prices they demonstrate complex correlations. We shall assume under the random walk hypothesis that the variables are random and independent.

Log price series' of assets have been traditionally thought to be well satisfied by the random walk hypothesis for a number of reasons. Firstly, as a price series has no fixed level it can be thought of as being non-stationary. A stationary series is a series that is invariant under time shift. Secondly, a price series if differenced once (as done to get returns), will become stationary, the definition of a unit-root series. A non-stationary unit-root series such as this can be best described by a random walk which is neither mean reverting nor predictable.

An important theory accompanying the random walk hypothesis for stock price is the concept of market efficiency.

3.1.1 Market Efficiency

The random walk hypothesis is based on the assumption of market efficiency. This assumes that the present stock price contains all current information available to forecast the future price and it is the only factor that has an effect in the future stock price. As new information enters the market any unbalanced states are instantaneously discovered and quickly amended by a correct change in the market place. Therefore under the market efficiency hypothesis prices reflect rapidly all available information in a completely unbiased way. This essentially means that the present stock value is all you need to determine the future stock price. According to this perspective a look back at historical prices, known as chart analysis, is worthless.

As this original definition of market efficiency was demanding and unre-

alistic in its totality it was suggested by Fama (1970) that market efficiency could in fact be subdivided into three categories. These three forms of market efficiency were differentiated by the level of strictness to which they followed the base definition of market efficiency. The three forms are:

1. Weak Market Efficiency

2. Semi-Strong Market Efficiency

3. Strong Market Efficiency

Weak Form Market Efficiency assumes that only past price data is considered when valuing a stock. This form of market efficiency rules out any manner of future price prediction based on anything other than past stock price data. It is assumed that the stock follows a completely random walk in which successive changes have zero correlation. This rules out any possible methods of prediction being consistently meaningful.

Semi-Strong Form Market Efficiency assumes that as well as stock price data information also all other publically available information is assimilated in the present stock price. This publically available information can include trading data such as volume data and fundamental data such as sales forecasts.

Strong Form Market Efficiency is the most stringent of the three forms. It is based on the assumption that all information available at present, both publically and privately, is considered and reflected in the present market. This is hard to verify in reality. Studies into strong form market efficiency

have revealed that insiders can and regularly do make exceptional returns thus creating an inefficient market at strong form level.

3.1.2 Definition of a Random Walk

Mathematically speaking, a time series s_t is a random-walk if

$$s_t = s_{t-1} + a_t$$

where s_t is the logarithm of the stock price at time t and $\{a_t\}$ is a white noise series.

A white noise series is a series of real numbers $\{w_i\}$ that are independent and identically distributed random variables that are symmetrical around a finite mean with a finite variance. If each w_i is normally distributed, then a white noise process is a Gaussian white noise process. It is a stationary series and it has the following properties:

$$Mean = E[W_i] = 0 \qquad \text{(by definition of the series)}$$
$$Cov(W_i, W_j) = 0 \qquad \text{(by independence assumption)}$$

3.1.3 Applying the Hypothesis to Financial Series Data

It is assumed that stock returns are independent random variables and if the time intervals are equal in length then the returns can be taken to be identically distributed also. That is to say if $S(t_i)$ denotes the stock price at

time t_t then the simple returns

$$\frac{S(t_1)}{S(t_0)}, \frac{S(t_2)}{S(t_1)}, \ldots, \frac{S(t_n)}{S(t_{n-1})}$$

are independent and identically distributed random variables. It can be then shown that stock prices follow a geometric random walk as follows

$$\begin{aligned} \frac{S(t_n)}{S(t_0)} &= \frac{S(t_n)}{S(t_{n-1})}.\frac{S(t_{n-1})}{S(t_{n-2})} \cdots \frac{S(t_2)}{S(t_1)}.\frac{S(t_1)}{S(t_0)} \\ &= Y(t_n).Y(t_{n-1}) \ldots Y(t_2).Y(t_1) \end{aligned}$$

Where $Y(t_i) = \frac{S(t_i)}{S(t_{i-1})}$ (ie. the simple return at time t_i)

Therefore,

$$S(t_n) = S(t_0) \prod_{i=1}^{n} Y(t_i)$$

which is a geometric random walk. Taking the natural logarithm of both sides we get;

$$\begin{aligned} lnS(t_n) &= ln\left[S(t_0) \prod_{i=1}^{n} Y(t_i)\right] \\ &= lnS(t_0) + ln\left[\prod_{i=1}^{n} Y(t_i)\right] \\ &= lnS(t_0) + \sum_{i=1}^{n} lnY(t_i) \end{aligned}$$

It can be seen from this that the natural logarithm of the stock price is a random walk of the form

$$s_t = s_{t-1} + a_t$$

As stated earlier it is often assumed that these log returns follow a normal distribution.

3.2 Testing for Normality

3.2.1 Overview of Normality in Returns

The first stylized fact about financial asset returns that will be tested is the assumption that returns are normally distributed. Essentially we wish to accept or reject the ideology that returns are independent and identically distributed (i.i.d) random variables, symmetric about a centre value, the mean with a finite variance. Under this normality assumption only two fixed parameters, the mean and the variance are needed to fully describe its distribution.

If the log values of a random variable are i.i.d as normal then it follows that the raw values of the random variable are independent and identically log-normally distributed. The mean and variance of the simple returns can then be found using the formulae (Tsay, 2002):

$$E(r_t) = e^{\mu + \frac{\sigma^2}{2}} - 1$$
$$Var(r_t) = e^{(2\mu+\sigma^2)}(e^{\sigma^2} - 1)$$

where μ and σ^2 are the mean and variance of the normally distributed log returns respectively. Alternatively if the simple returns are known to be log normally distributed with mean m_1 variance m_2 then the mean and the variance of the log returns are given by:

$$E(r_t) \quad = \quad ln\left[\frac{m_1+1}{\sqrt{1+\frac{m_2}{(1+m_1)^2}}}\right]$$

$$Var(r_t) = ln\left[1 + \frac{m_2}{(1+m_1)^2}\right]$$

Due to huge amount of past research into the normal, or Gaussian, distribution there are a vast number of important procedures such as t-tests, analysis of variance and tests for regression coefficients that are based on the underlying assumption that the sampled data came from a normal distribution. Therefore before applying such procedures it is imperative to test the robustness of the normality assumption on the sampled data.

As the normal distribution is, as its name suggests, the most common and desirable distribution for statisticians to use it is often assumed that a data set is normally distribution until proven otherwise. Due to its advantageous statistical properties it takes strong evidence to dismiss a normality assumption completely. Even if a set of data is not normally distributed it is often attempted to bend the rules or apply transformations so that useful procedures that have been defined for the normal distribution can be applied.

The reason for trying to accept or reject the assumption of the data following a normal distribution, or any other distribution for that matter, is simple and logical. If we wish to make conjectures particular to a given distribution we need to first verify that the data actually follows the given distribution otherwise the resulting conclusions may, and probably will, be untrue. If decisions are to be made based on these untrue conclusions then in turn these decisions will be unsound. It can be seen that a false base assumption can have a negative impact on future conclusions and decisions. Conversely a well grounded initial assumption can allow other conclusions to

be drawn on a good basis. Also when dealing with figures confidence levels are often required and a seemly acceptable confidence level can be completely wrong if the underlying assumption is even slightly incorrect.

Due to the simple yet specific characteristics of the normal distribution, such as symmetry, it is not an overly strenuous task to discount data that follows a completely different distribution. A more arduous task is proving that the data is actually normal or close enough to normal that Gaussian assumptions can be well grounded. Types of data sets that the normal distribution may not be acceptable to model might include:

- A bounded data set: The Gaussian distribution is unbounded on the entire real axis meaning it can take any real value along the x-axis. Therefore it is almost certainly acceptable not a good fit for bounded data.

- Asymmetric data: One of the most obvious properties of the Gaussian is that it is symmetric therefore it is not suitable to model left- or right-skewed data.

As explained earlier many financial concepts have been built on the traditional concept that return values of assets are normally distributed with fixed mean and variance. This assumption makes returns look very appeasing for statisticians but unfortunately a number of characteristics of log returns disagree with the above points and that indicate the data may not be normal (Tsay 2002):

- The lower bound of a log return 0 as it cannot take on any values less than or equal to this whereas a normally distributed random variable can assume any value on the real line giving it no lower bound.

- Empirical asset returns tend to display positive excess kurtosis which is not a characteristic of the normal distribution.

Nevertheless for fear of losing the very useful assumption of normality, we shall examine the distributional properties of the data with more rigorous tests of normality.

3.2.2 Exploratory Data Analysis

First to get a feel for the data's distributional properties we shall complete some exploratory data analysis. As stated by Henry C. Thode in his book 'Testing for Normality' (Thode, 2002) the methods of testing for normality, both graphical and numerical are almost infinite. To carry out the graphical exploratory analysis on the data we will use histograms, Quantile-Quantile (Q-Q) plots, mean excess plots and kernel density estimation plots. As part of this a table of summary statistics of the data will be useful to compare the individual return distributions.

Histograms

Histograms are a graphical representation of the frequency distribution of a data set. They provide an insight into skewness, tail behaviour, outliers and multimodality. The basic procedure of constructing a histogram consists of

dividing the interval covered by the data set into equal length sub-intervals known as bins. The frequency of data values falling into each bin determines the height of the bin such that the more values falling within the bin interval the higher the bin will be. Therefore it can be seen that histograms give a rough estimate to the actual distribution.

Histograms are a good starting point for analysing the shape and location of the data distribution but some of its properties such as being non-smooth, depending on the end points of the bins and depending on the choice of bin width can be unsatisfactory. The shape of the histogram will be particularly influenced by the choice of end points. This can lead to the appearance of multimodality even if it is not actually a property of the data distribution.

Looking at all the histograms of the monthly log returns (Figures 3.1 - 3.4) it is immediately obvious that they are not too dissimilar to the normal distribution. They are all unimodal distributions that decay in frequency either side of this mode. It is evident in all of the graphs of returns that there are extreme values present at the far left of the histograms. The graphs are reasonably symmetric were it not for the presence of these outliers. The affect outliers have on the distribution is to cause it to become skewed in the direction in which they lie. Therefore we suspect some negative skewness in the returns.

With the log returns if they display normality we expect to see a symmetrical graph. We observe that the log returns look close to normal and centred about zero in general. The log returns of Boeing appear almost perfectly symmetrical and give a strong indication of being normal. The log

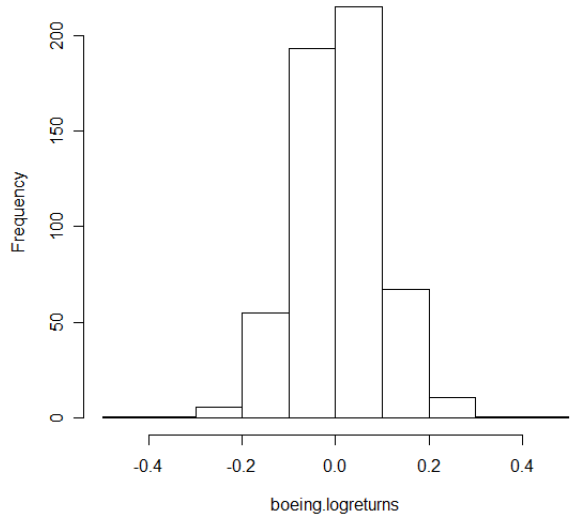

Figure 3.1: Histogram of Boeing Monthly Log Returns

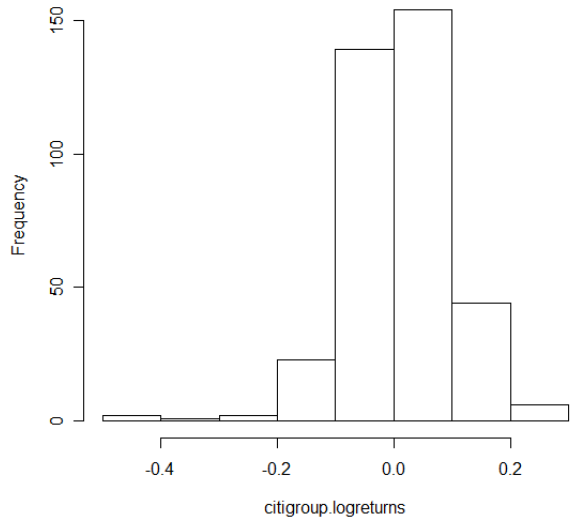

Figure 3.2: Histogram of Citigroup Monthly Log Returns

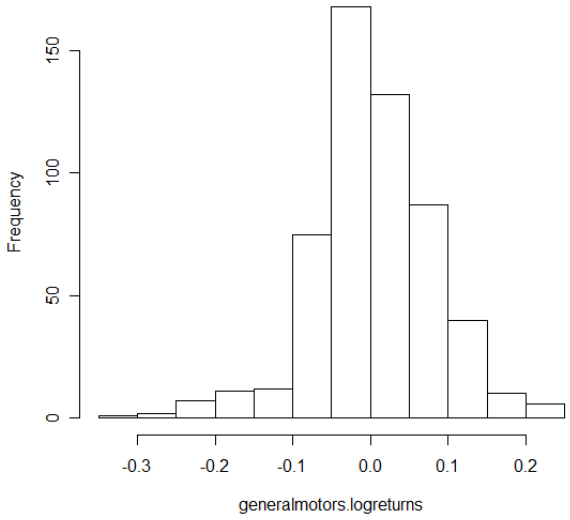

Figure 3.3: Histogram of General Motors Monthly Log Returns

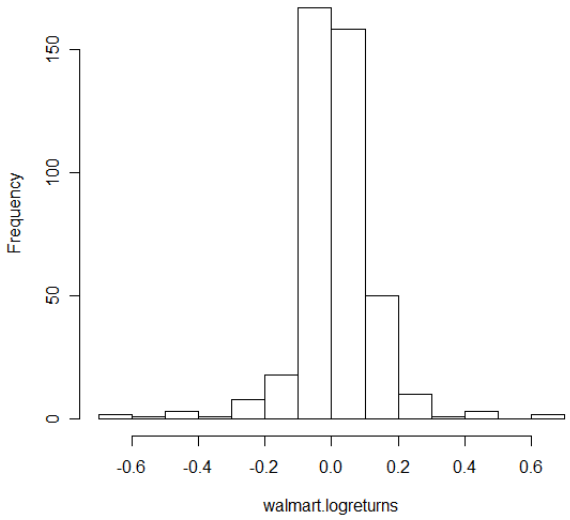

Figure 3.4: Histogram of Wal-Mart Monthly Log Returns

returns of Wal-Mart also show symmetry about the mean but the histogram exhibits a much quicker decay either side of the high peak at the mode. It also displays elongated tails strongly hinting at excess kurtosis. The Citigroup, General Motors and Intel histograms are less symmetrical than those of Boeing and Wal-Mart demonstrating extreme values on their lower sides. This characteristic, along with the fact that most of the mass of these distributions is located in the vicinity of their modes gives an indication of negatively skewed distributions for the monthly returns for these stocks.

The presence of symmetrical stock returns for the other companies means we can't conclusively accept skewness as a characteristic of monthly returns in general. The histogram of monthly returns of the Boeing stock price is unimodal with a peak close to zero. The evidence viewed in the histogram of the Boeing returns gives the impression that a normality assumption might be correct at a monthly interval of return.

As regards the other companies' returns, the presence of outliers and positive skewness is evident. The outliers are located at -0.55, -0.425, -0.85 respectively for Citigroup, General Motors and Intel. These are located to the extreme left relative to the mode at approximately 0 in each case. It is also important to notice that the mode is slightly positive in four out of five stocks with General Motors being the only one displaying a negative mode.

As seen from the table of summary statistics in Figures 3.5 and 3.6 all the means of the stocks returns are positive adhering to the expectation that stocks increase on average. This is logical as otherwise there would be no reason to invest in them. Also again in relation to the properties

Company	Minimum	Maximum	Mean	Median
Boeing	-0.42419	0.417384	0.010352	0.009756
Citigroup	-0.439833	0.2948	0.009281	0.010695
General motors	-0.349999	0.23144	0.004496	0
Intel	-0.588738	0.388658	0.016836	0.018965
Wal Mart	-0.693147	0.693147	0.016172	0.005781
Dow Jones	-0.366737	0.305704	0.004171	0.008704

Figure 3.5: Table of Statistics for Monthly Data(1)

Company	Variance	Stdev	Skewness	Excess Kurtosis
Boeing	0.008879	0.094228	-0.043332	1.639829
Citigroup	0.007438	0.086244	-0.629543	3.595866
General motors	0.006068	0.077895	-0.297824	1.616722
Intel	0.01549	0.12446	-0.759806	2.295017
Wal Mart	0.014706	0.121268	-0.397279	10.874098
Dow Jones	0.002924	0.054072	-0.81166	7.254719

Figure 3.6: Table of Statistics for Monthly Data(2)

of skewness it can be seen from the table that all the minima are further from their respective means than the corresponding maxima. That is the absolute difference of the minimum of the return is greater than that of the maximum. This gives an indication of skewness across all monthly returns. We deduce that large negative returns are much more probable than high positive returns. A stock value can decrease substantially in a very short time interval whereas it needs much more time to recover its value.

Further evidence from the table of statistics of the log returns comes to support the claim of skewness and kurtosis in the last 2 columns where there is negative skewness and excess kurtosis resulting from all the empirical returns distributions.

Kernel Density Estimation

As remarked earlier some properties of histograms can be unsatisfactory. These include being non-smooth, depending on the end points of the bins and depending on the choice of bin width. The first two aspects can be alleviated with use of kernel density estimation. Kernel density estimation is a non-parametric technique for estimating the density of data. This method centres a function called a kernel at each data point instead of fixing the end points. This negates the dependence on the end points that is present in histograms.

The density function acquired from this method is defined as

$$f_b(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right)$$

where K is the selected kernel density, b is the bandwidth chosen for the kernel function. Due to the division of the sum by nb the function integrates to 1, that is

$$\int_{-\infty}^{\infty} f_b(x)dx = 1$$

K is a non-negative function chosen such that it is the density of a centred random variable such that

$$\int_{-\infty}^{\infty} Kdx = 1 \qquad and \qquad \int_{-\infty}^{\infty} xKdx = 1$$

Choice of b is important as a smaller b gives a lower bias but a higher b gives a lower variance. The weakness of the method is the arbitrary choice of K and b.

An example of a kernel that can be taken is the box kernel. Using this method centres a block at each data point creating a graph similar to a histogram but less blocky. The problem with this method is that due to the kernel being discontinuous the density function built from it is also. To get a smooth density estimate a smooth kernel is required.

A smooth, continuous kernel will smoothen out the appearance of the graph and eliminate the dependence on the choice of bin endpoints. However it is not possible to eradicate the reliance on bandwidth. If a kernel with a bandwidth that is too small is chosen there is a risk that the estimated density will be under-smoothened resulting in more local modes than there should be. Alternatively if the bandwidth of the chosen kernel is too large an over-smoothened density estimate which will hide many of the feature of the actual underlying density. There are a number of methods that can be used

to choose the optimal bandwidth. One such method takes the bandwidth as the argument that minimises the Asymptotic Mean Integrated Squared Error (AMISE). This method recovers all the important features of the density and maintains smoothness (Duong, 2001).

As it was the log returns that showed the greatest potential to be normally distributed at a monthly frequency the kernel density estimation method will be concentrated on assessing these. R was used to superimpose the density estimates onto the relevant histogram and a 'rug' of the data points was created beneath the X-axis showing the position of each data point. The underlying kernel we have chosen for all the estimation procedures is the Gaussian. It can be seen that the densities follow approximately the same pattern as the histograms predicted although some have higher and sharper peaks than the histograms exhibit.

The Wal-Mart Gaussian kernel density estimate shown below in Figure 3.7 gives an example of how the histogram can miss out on some features of the density. The density estimate shows at very steep peak at 0 and a sharp declining slope either side. The bulk of its density is located in the centre with two thin elongated tails either side. This is a feature that leads to excess kurtosis.

The Boeing estimate in Figure 3.8 has a more normal bell-shape to it. It has gentler gradients either side of the peak. The Citigroup (Figure 3.9), General Motors and Intel density estimates show a much more skewed density than those of the other two stocks' log returns. All three are negatively skewed, with large pointed peaks and a long tail extending to the left.

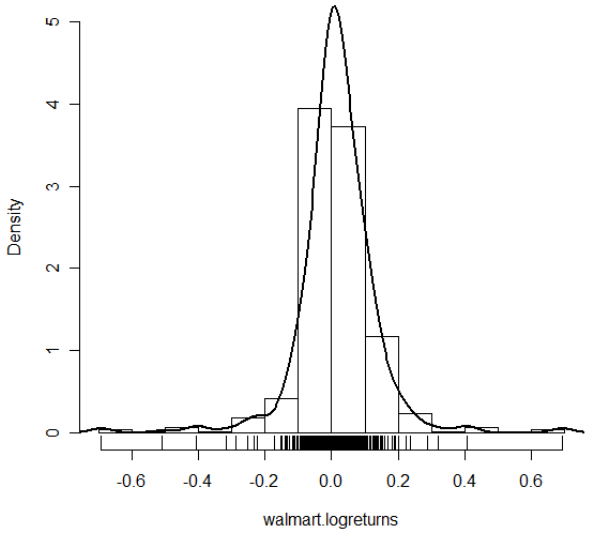

Figure 3.7: Kernel Estimate of Monthly Returns of Density for Wal-Mart

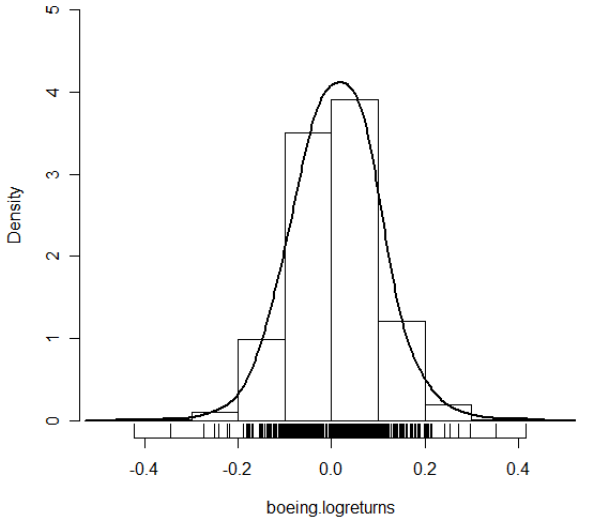

Figure 3.8: Kernel Estimate of Monthly Returns of Density for Boeing

Quantile-Quantile Plots

Histograms and kernel density estimators are only the first step in trying to get an overall idea about how the data looks and to investigate if it behaves like any other known distribution. Another informal graphic diagnostic that is popularly used in this process is the Quantile-Quantile or Q-Q plot. The Q-Q plot is a probability plot which is a graphic in which the empirical order statistics on the Y-axis are compared to expected values of some theoretical order statistics located on the X-axis.

As we are testing for normality the Q-Q plots will plot the standardized empirical quantiles of the observed data against the quantile of a standard normal random variable. The standardized returns are used under the as-

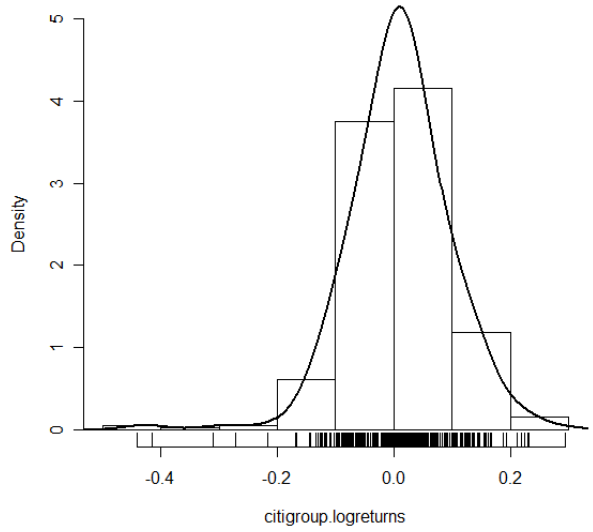

Figure 3.9: Kernel Estimate of Monthly Returns of Density for Citigroup

sumption that the data is reasonably normal and they are utilized to have the X-axis and Y-axis scaled the same. The data can be standardized using

$$z = \frac{x - \mu}{\sigma}$$

A quantile is a figure that a certain percentage, or fraction, of the data lies below in the distribution. More specifically, the q-quantile of a random variable X is any value x such that

$$\begin{aligned} Pr(X \leq x) &= q \\ &or \\ F(x) &= q \end{aligned}$$

where F is the cumulative density function. Taking the inverse of the equation gives you the quantile function defined as follows

$$x = F^{-1}(q)$$

If a random variable is not continuous quantiles may not be unique or may not exist at all. The 0.25-, 0.50-, and 0.75-quantiles are commonly known as the first, second and third quartiles. The 0.01-, 0.02-, 0.03,...-quantiles are called the first, second, third, ... percentiles. The Q-Q plot is then the plot of these percentiles of the empirical distribution against the theoretical distribution that you wish to compare it to.

Assuming that the data is normally distributed we will expect to observe a linear plot except for some random fluctuations in the data. To aid with determining this, a line with a slope of 1 is superimposed on the Q-Q plot.

Any systematic deviation from this line will indicate that the data is non-normal. Due to its nature, that is requiring opinion to determine the severity of a deviation, it is more a judgemental method of assessing normality than a formal one.

For a heavy-tailed distribution one would expect to see the upper tails of the Q-Q plot turning upwards and the lower tails bending downwards. Alternatively for a short-tailed distribution it would be expected to observe an S-shape with the lower tails bending upwards and the upper tail curving downwards. Another property of a distribution noticeable in its Q-Q normal plot is symmetry. A symmetric distribution will typically produce a symmetric Q-Q plot that is linear in the centre of the data. When the data is standardized and Q-Q plotted normal parameters can be estimated from the plot using regression.

Looking at the Q-Q plot of the Boeing log returns we observe that it is linear near the middle between -2 and 2 and it looks almost symmetrical about 0 also. There seem to be systematic deviations from the 45° line at either end of the plot in comparison with the simulated data Q-Q plot in which the end points are much closer to the line. For the Citigroup data shown below there is also a very strong departure from the line at the negative end of the plot suggesting it is definitely not behaving like the normal distribution up in the lower tail. It takes a value of -4 in the standardized empirical log returns to cover the same fraction of data that a quantile of -3 covers in a normal distribution. This is a strong indication of a heavy left tail. It is worth noting also that the upper end of the Citigroup Q-Q plot

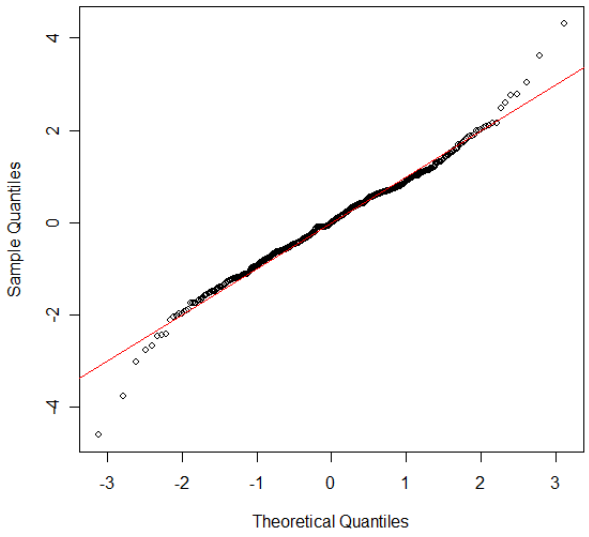

Figure 3.10: Q-Q plot of Standardized Monthly Log Returns for Boeing

lies almost completely on the superimposed line therefore it follows a much more normal behaviour than the lower tail.

As a comparison a sample set of randomly generated data from an i.i.d normal distribution was also plotted against the normal in a Q-Q plot and this is shown just below the Citigroup Q-Q plot. It demonstrates the linear plot that should be observed for a sample of the same size as the Citigroup log returns data if the data were normal.

The General Motors and Intel Q-Q plots (Figures 3.13 and 3.14) are similar to the Citigroup plot in that they show linearity about the centre with serious deviation from the line at the negative end of the plot. A difference between the General Motors and Intel plot is observed in the upper tails

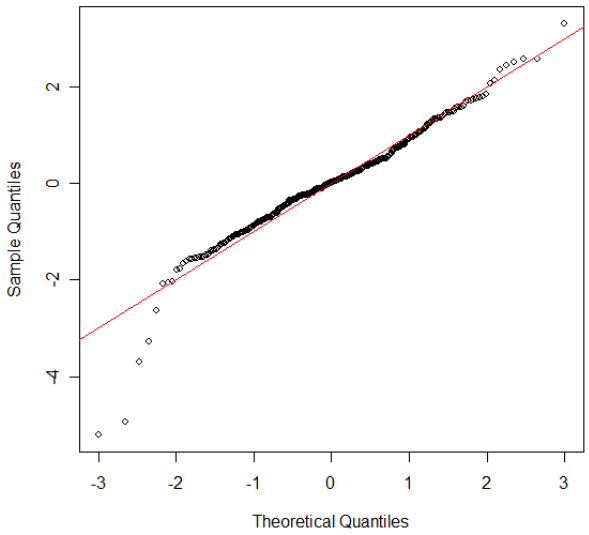

Figure 3.11: Q-Q plot of Standardized Monthly Log Returns for Citigroup

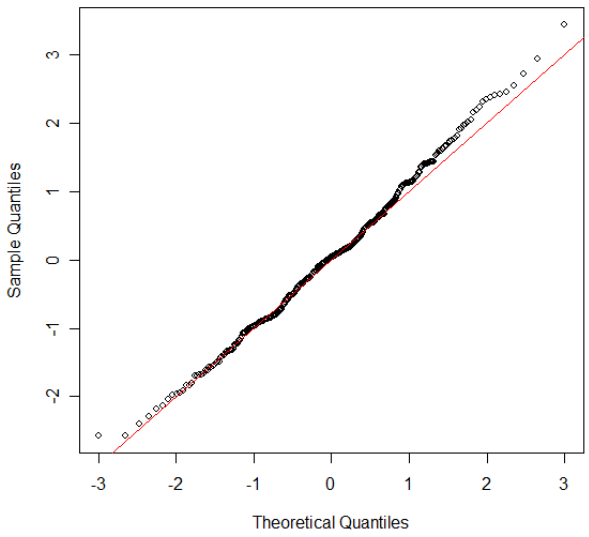

Figure 3.12: Q-Q plot of Simulated Normal Data

where the General Motors values are almost perfectly situated on the line while the Intel points actually move away from the line in an downward direction before returning to it again. This former characteristic might suggest short tailed tendencies in the upper tail of Intel.

The Q-Q plot of the Wal-Mart log returns in Figure 3.14 is almost perfectly symmetrical and in fact it is the most symmetrical of all five stocks. Even though it is almost certain that the underlying log returns distribution is symmetrical the huge deviations of the ends of the plot from the 45 ° line suggest large kurtosis which will result in much longer than normal tails of the density. This feature has already been observed from the corresponding histogram.

The Q-Q plot shown below is for the monthly Dow Jones standardized log returns. Again it displays symmetry but there are more points deviating from the line in the lower side of the graph than in the upper side giving an indication of a heavier negative tail than upper one.

Mean Excess Plots

This repeating feature of deviation of the tails away from the 45 ° line induces interest in the activity of the returns at the tails. A first step into investigating the tails might be to use a mean excess function to describe them. A mean excess function describes the expected surpassing of a threshold provided that exceedance of this threshold has taken place. As the mean excess plot is only advantageous for looking at the upper tail of a distribution we will define the loss distribution as the negative of the return distribution to

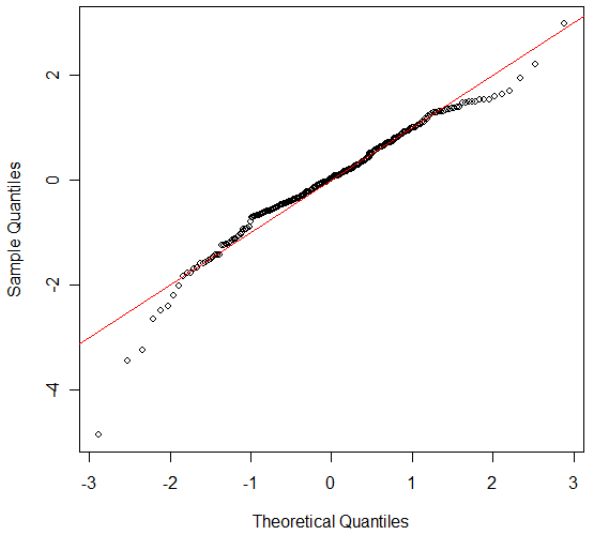

Figure 3.13: Q-Q plot of Standardized Monthly Log Returns for Intel

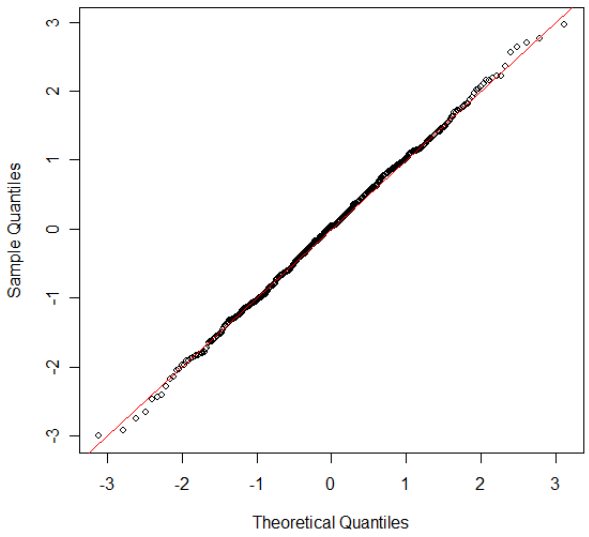

Figure 3.14: Q-Q plot of General Motors Standardized Monthly Log Returns

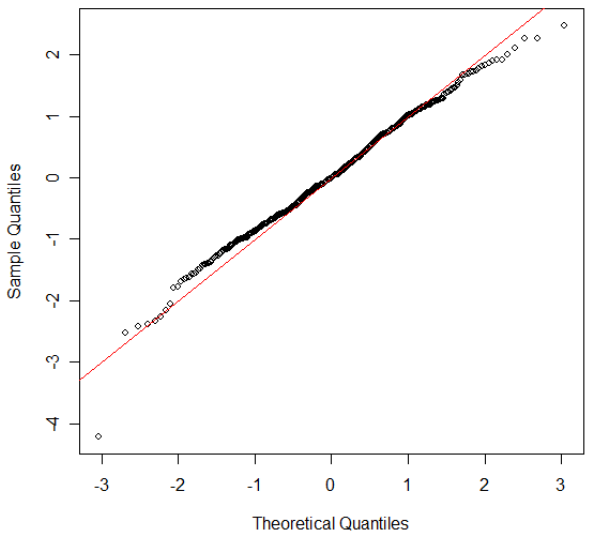

Figure 3.15: Q-Q plot of Wal-Mart Standardized Monthly Log Returns

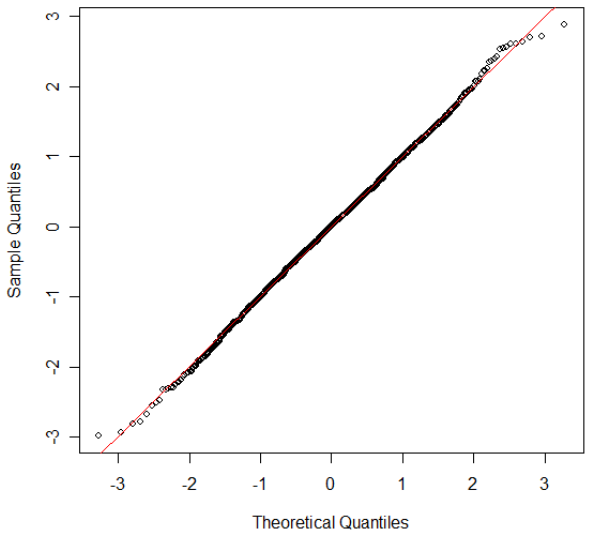

Figure 3.16: Q-Q plot of Dow Jones Standardized Monthly Log Returns

allow use to use it on the lower tail. Fundamentally the mean excess function is:

$$e(u) = E[X - u|X > u]$$

where $e(u)$ is the mean excess function and u is the variable threshold of the function.

For a sample of size n the mean excess function is more explicitly defined by

$$e_n(u) = \frac{\sum_{i=1}^{n}(X(i) - u)}{\sum_{i=1}^{n} 1_{X(i)>u}}$$

where $X(i) - u$ is only valid when it is positive, i.e. when $X(i)$ is an observation that exceeds the threshold. This essentially defines the sample mean excess function as the sum of excesses over the threshold u divided by the number of data points which exceed the same threshold (McNeil 1999). The sample mean excess function is the empirical estimator of the mean excess function corresponding to that distribution the sample is taken from. Thus for a large sample size the sample mean excess function almost coincides with the mean excess function corresponding to the underlying distribution. An important property of the mean excess function is that any continuous distribution function is uniquely determined by its own mean excess function. A plot of the function is then simply the paired values of the threshold u and the mean excess function value $e_n(u)$ between the first and n^{th} order statistics of the sample, i.e.

$$\{(u, e_n(u)), X_{1:n} and X_{n:n}\}$$

where $X_{1:n}$ and $X_{n:n}$ are the first and n^{th} order statistics.

The conclusions drawn from the plot should be based on the gradient of the line. The trend of a positive gradient indicates heavy tailed behaviour. In particular if the empirical mean excess plot follows a reasonably straight line with an upward trend above a particular threshold this is an indication that the excess over this threshold follow a generalised Pareto distribution defined with a positive shape parameter. A downward trend in the plotted line indicates thin tailed behaviour while a line with a gradient of 0 implies an exponential tail.

It is important to realise when reading the graph that the upper points plotted are calculated as the average of only a handful of extreme excesses so they may be erratic. This is a problem when examining linearity in the graph. For a nicer looking plot some of these upper points are excluded. In the mean excess plot function included in the evir package for R, meplot(), the default number of data points omitted is 3. It can be set to the required number of omissions with use of the parameter specification 'omit=', e.g. meplot(x, omit=1) where x is the data vector to be plotted. As most of the data hint at heavy-tails we will use the mean excess function to plot both the lower tails i.e. losses and the upper tail which contains the profits.

After examining the mean excess plots generated by R there is a common appearance to the plots. All the plots have strong negative slope until at least a threshold of -0.1. The line up to this point is very smooth with no sudden jumps or drops. This is mainly due to the large number of data points that exceed the lower thresholds. Across all the assets the plot has downward trend until close to 0 and then as points above the threshold become fewer

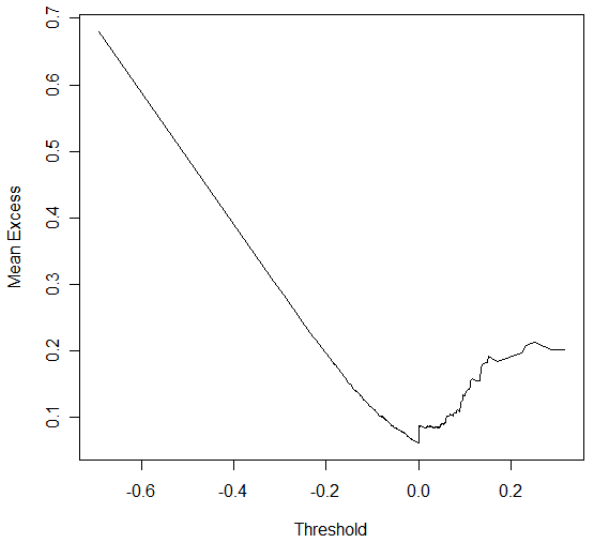

Figure 3.17: Mean Excess Plot of the Monthly Wal-Mart Losses

the plot becomes more jagged and erratic especially for the last number of points on each plot.

The Wal-Mart mean excess plots for its tails show the most positive linear trend at the upper end of the graphs. The plot is less erratic than the other stocks but this was partially due to the scale of the Y-axis which has a much greater mean of excess than the other stocks. The positive gradient on the positive side of the plot indicates that in the upper tail the distribution may follow a generalised Pareto distribution with a positive slope parameter. It can be seen that the upper part of the mean excess of the loss function is more boldly linear than the plot for the regular return.

Although the other stocks do not show as clear a linear positive slope in

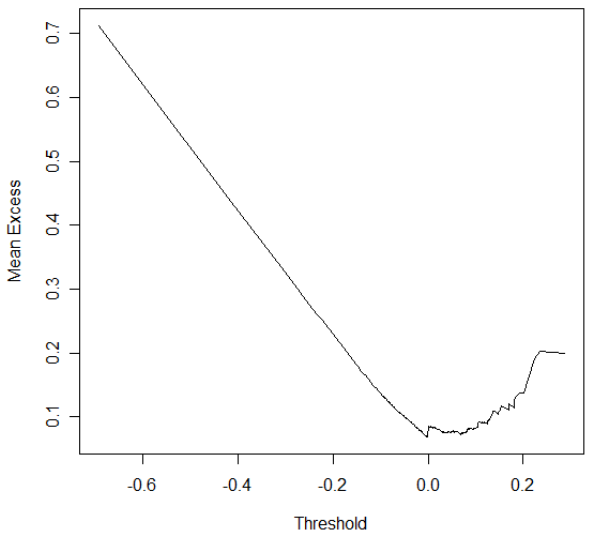

Figure 3.18: Mean Excess Plot of the Monthly Wal-Mart Log Returns

the upper part of their mean excess plots the upward trend is still evident. The mean excess plot of the Citigroup log returns below shows some very erratic movement in its upper end. It has a sharp rise between the thresholds of 0.1 and 0.16 and then the line fluctuates greatly after this. This characteristic will need to be queried when studying a greater frequency of observation.

Looking at the mean excess plot for the Dow Jones monthly losses we see that it is less erratic than the single stock log returns, in general. Again it shows a substantial negative slope until the threshold of 0 followed by a positive gradient. This leads to the speculation that again a generalised Pareto distribution may suit modelling the tail of the losses. As the number

of extremes becomes more influential the plot becomes more erratic and loses its linearity. Due to the sample size the effects can be ignored.

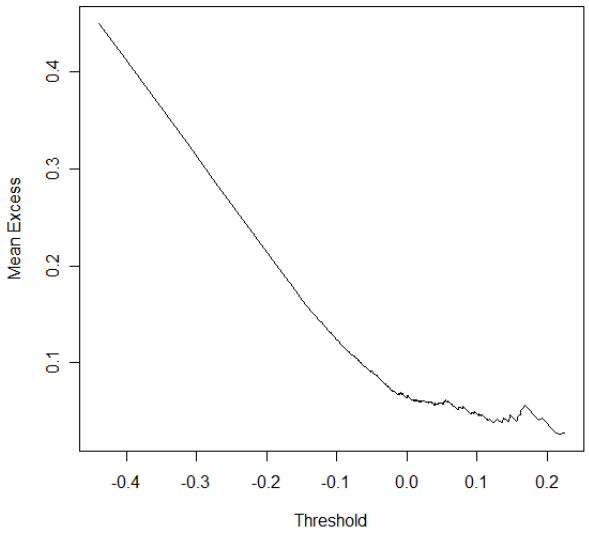

Figure 3.19: Mean Excess Plot of Citigroup Monthly Log Returns

So far although we have seen some characteristics of the normal distribution shine through the graphs and some symmetry in the Boeing and Dow Jones monthly log returns over all the assumption of normality is hard to accept or reject with mixed results coming from the graphs overall. Putting the returns through some statistical tests will give a more subjective opinion of the normality of the stocks. If these fail to draw conclusive deductions we may need to look at the returns at a greater frequency of observation.

3.2.3 Statistical Tests of Normality

Probability plots can be sensitive to random observations in the data and exclusive reliance on them can lead to erroneous conclusions as demonstrated by D'Agostino, 1986. Therefore it is crucial to have more objective methods to validate assumptions made about distributions. Statistical tests can be used as more deterministic methods of assessing normality.

The skewness and kurtosis are important characteristics in determining whether data follows a normal distribution or not. Many commonly used statistical tests are based on the expected values for the kurtosis and skewness of a distribution therefore it is important to define these concepts before looking at the tests that use them.

Skewness and Kurtosis

As defined earlier skewness is the third standardized moment and it measures the lack of symmetry in a distribution. The skewness of the Gaussian distribution is zero and any other symmetrical distribution should have a skewness of close to zero also. Negative values of skewness indicate that the data is skewed left, or negatively skewed while positive values tell the data being skewed right, or positively skewed.

The skewness of a distribution is given by:

$$\gamma_2 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

where μ_i the i^{th} central moment. For a sample of a distribution an estimate

of the skewness is given by

$$b_1 = \frac{k_3}{k_2^{\frac{3}{2}}}$$

where the k_i are the unbiased estimator of the i_{th} cumulant. There are a number of definitions of this estimate for sample skewness but the one that is used in the R is

$$b_1 = \frac{\sum_{i=1}^{N}(y(i) - \bar{y})}{(N-1)s^2}$$

If the data is skewed from the normal we would expect to find that the skewness is not close to 0. The kurtosis is the fourth standardized moment of the distribution and for the normal distribution it is exactly three. There are mixed ideas about the exact characteristics that the kurtosis describes (Thode, 2002). Thode says that the kurtosis describes the density and dispersion at and around $\mu \pm \sigma$, areas he calls the shoulders of the distribution and that with high kurtosis figures there is less mass at these shoulders and so it is concentrated at the centre and/or the tails of the distribution.

Therefore for reasonably well behaved distributions a kurtosis figure higher than 3 it indicates that the distribution has heavy tails and peaks close to its mean. If less than 3 the kurtosis tells us that the sample data has a flatter distribution than the normal. Although not getting a kurtosis of 3 indicates that data is not normal getting a kurtosis of 3 does not mean that the data is normal. As shown by D'Agostino and Lee (1977) symmetrical distributions with the same kurtosis can have very different estimation efficiencies.

Kurtosis of a distribution is defined as

$$\gamma_2 = \frac{\mu_4}{\mu_2^2}$$

where μ_i are as described above for skewness. The sample kurtosis is then given by

$$b_2 = \frac{k_4}{k_2^2}$$

where k_i is defined as above for the sample skewness. On the next page are bar charts used to compare the skewness and excess kurtosis values of the monthly log returns for all the companies and the Dow Jones Industrial Average. It is observable that the kurtosis values are all positive, a clear indication of heavy tails. Due to the nature of kurtosis it does not give information as to whether the distribution has one or two heavy tails and if it only has one which end is it at. Using the bar charts of skewness we can get a clearer view as to the dispersion of the mass in the distribution. All the data have a sizeable negative skewness value, except for Boeing for whom we had earlier noted to be reasonably symmetrical at a monthly frequency. This supports the concept of heavy lower tails.

The most striking feature of the bar charts for the log returns is the huge difference between the Wal-Mart kurtosis and the kurtosis of the other stocks log returns. Only the kurtosis of the Dow Jones log returns is anywhere near that of Wal-Mart. From the bar charts it can be estimated that the Boeing is closest to normal having the lowest absolute scores for skewness and kurtosis overall. The distributions of the monthly log returns for Citigroup and Intel are quite similar suggesting noticeably heavier negative tails indicating the presence of a number of extreme losses. While General Motors is both less skewed and has less kurtosis that the two previously mentioned it still

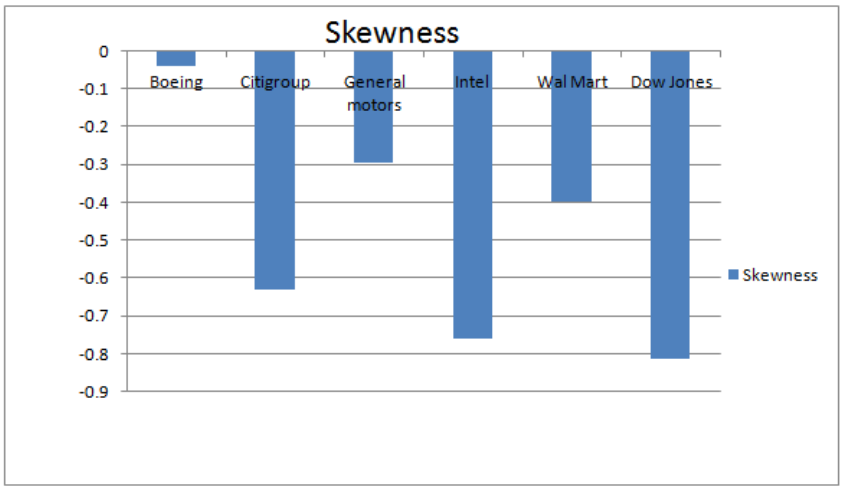

Figure 3.20: Bar chart of Monthly Skewness

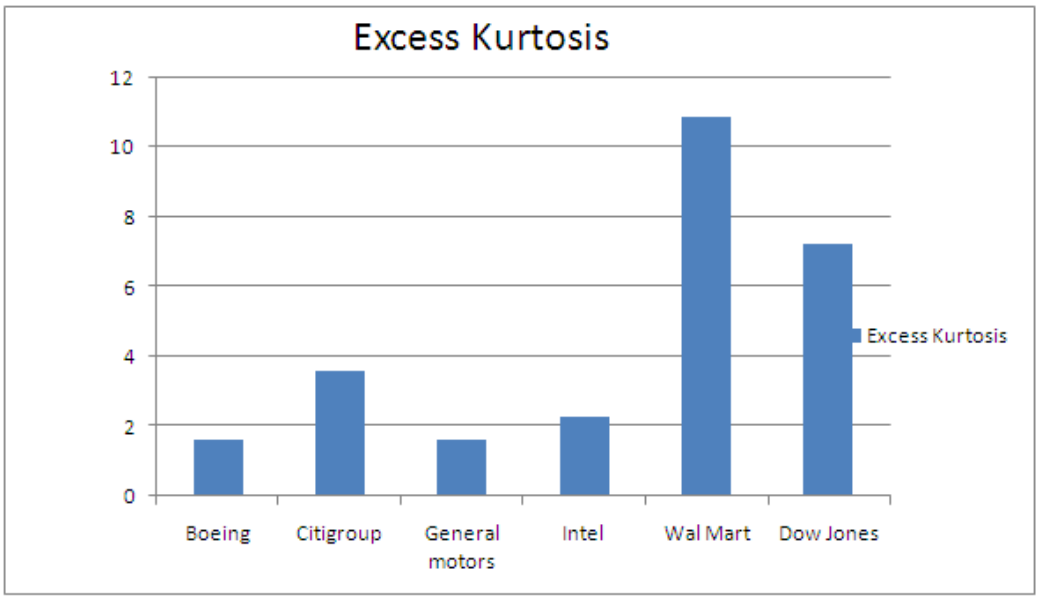

Figure 3.21: Bar chart of Monthly kurtosis

displays the same characteristics just with a less severe deviation from the normal. The Dow Jones log returns show considerable skewness and kurtosis which means it is heavily skewed and heavy-tailed, or leptokurtic. The expectation follows that at this frequency its returns will be completely rejected as being normal by the normality tests.

The tests we will use to examine the log returns (and inclusively the losses) against a null hypothesis of normality will be the Jarque-Bera test, the Shapiro-Wilk test and the Kolmogorov-Smirnov test.

The Jarque-Bera test

The Jarque-Bera test examines the skewness and kurtosis of the data sample to see if it matches that of the normal distribution. It is one of the simplest and very likely the most commonly used procedure for testing normality of financial time series returns. It offers a joint test of the null hypothesis of normality in that the sample skewness equals zero and the sample kurtosis equals three.

The Jarque-Bera test statistic is calculated from

$$JB = \frac{n}{6}\left(b_1^2 + \frac{(b_2 - 3)^2}{4}\right)$$

where n is the same size, b_1 is the sample skewness and b_2 is the sample kurtosis as previously defined.

The null hypothesis is rejected if the test statistic exceeds some predefined critical value, which is taken in the asymptotic limit from the χ_2^2 distribution.

The Shapiro-Wilk test

The Shapiro-Wilk test is a test for the rank correlation between the empirical data and that of a sample from a normal distribution. It is a regression test that compares an estimate of the empirical standard deviation using a linear combination of the order statistics to the theoretical normal estimate. It concentrates on the slope of a plot of the order statistics versus the expected normal order statistics.

The Shapiro-Wilk test statistic is defined as

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{(x_i - \bar{x})^2}$$

where x_i are the order statistics from the empirical sample, $\bar{x}$ is the mean and a_i are appropriate constant values attained from the means and covariance matrix of the order statistics. These values for a_i can obtained from a statistical table but we shall let R do the hard work here.

Again if the statistical value is higher than a preordained critical value we fail to reject the null hypothesis and the normality assumption will stand.

The Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test investigates the null hypothesis that a sample belongs to a certain solution by looking at the maximum difference between the empirical and theoretical distribution functions. To be particular it concentrates on the difference between the empirical cumulative density function (ECDF) and the theoretical cumulative distribution. The ECDF is a step

function defined as

$$ECDF = \frac{n_i}{N}$$

where n_i is the number of points less than Y_i where Y_i the set of ordered data points and N is the total number of data points.

The Kolmogorov-Smirnov test is in fact not particular to testing for normality as it is a general goodness of fit test. It is important to be aware that it is not as powerful as more normal specific tests like the dependable Shapiro-Wilk test previous described. Problems with the Kolmogorov-Smirnov test are that it only applies to continuous distributions which have been fully defined in terms of their parameters and it is more sensitive to data located at the centre than at the tails.

Results of Statistical Normality Testing

Examining the results of performing these three tests should allow us to test the assumptions we have already suggested giving a much clearer and more solid concept of the normality of the returns. For each of the tests we will have a null hypothesis H_0 that says that the data comes from an i.i.d Gaussian distribution and an alternative hypothesis saying that it does not. For the significance levels we will use 0.01, 0.05 and 0.10. The null hypothesis will be rejected if the p-value is less than the significance level and we will fail to reject the null hypothesis otherwise. The table below summarizes the test results from applying the three tests described above on monthly data.

The left hand column under the test name is the test statistic represented

by a solitary letter, e.g. T for the Jarque-Bera test. The column next to this contains the corresponding p-value for each test statistic. It is this figure that is measured against the significance level. As the tests were carried out on a 32-bit platform due to memory restrictions any values less than 0.00000000000000022 are simply denoted as $< 2.2e^{-16}$. As the significance levels we are dealing with are nowhere near as minute as $2.2e^{-16}$ the implicit nature of these inequalities will not affect the conclusions.

The table in figure shows that under the Jarque-Bera test of the assets log returns we should reject the null hypothesis that they are normally distributed. The highest of these p-values is still very much smaller than even the lowest significance level. Looking at the Shapiro-Wilk and Kolmogorov-Smirnov test also the evidence seems to be against the assets log returns being similar to the normal distribution. According to the Kolmogorov-Smirnov test the normality assumption for each of the assets log returns should be emphatically discarded. Saying this it should be remembered that the Kolmogorov-Smirnov test is a goodness of fit test and therefore more general and not as powerful as the other two tests.

As a comparison to show the validity of the three tests simulated I.I.D normally distributed samples of the same size as each of the empirical data sets were also tested. As expected all the p-values are above the highest significance level of 10%.

After examining monthly log returns over the entire lifetimes of the assets we have come to the conclusion that the asset returns do display some normal characteristics for the most part. The histograms and kernel density

All Monthly Data	Jarque-Bera test		Shapiro-Wilk test		Kolmogorov-Smirno	
Log Returns	T	p-value	W	p-value	D	p-value
Boeing	63.1873	1.90E-14	0.9858	3.36E-05	0.4112	< 2.2e-16
Citigroup	228.5742	< 2.2e-16	0.9564	5.00E-09	0.421	< 2.2e-16
General Motors	69.4521	7.77E-16	0.9768	1.17E-07	0.4124	< 2.2e-16
Intel	83.4764	< 2.2e-16	0.9649	2.20E-16	0.3978	< 2.2e-16
Wal-Mart	2125.555	< 2.2e-16	0.8336	< 2.2e-16	0.4021	< 2.2e-16
Dow Jones	2200.385	< 2.2e-16	0.9061	< 2.2e-16	0.4422	< 2.2e-16

Figure 3.22: Results of Normality Tests on the Monthly Returns

Data sample is of the	Jarque-Bera test		Shapiro-Wilk test		K-Smirnov test	
same size as	T	p-value	W	p-value	D	p-value
Boeing	3.9551	0.1384	0.9954	0.1025	0.0371	0.4349
Citigroup	0.8296	0.6605	0.9973	0.8167	0.0401	0.591
General Motors	0.8069	0.668	0.9981	0.8234	0.0418	0.2903
Intel	1.1887	0.5519	0.9942	0.4328	0.0667	0.2027
Wal-Mart	2.5608	0.2779	0.995	0.1822	0.0464	0.3197
Dow Jones	2.0149	0.3652	0.9975	0.149	0.0257	0.5593

Figure 3.23: Results of Normality Tests on the Simulated Data

estimations for a monthly frequency displayed some symmetry in most of the stocks but from the tables there was negative skewness evident so we suspect they don't follow normal distributions.

An investigation of all the returns for a greater frequency will give more support to compound or refute these assumptions. The monthly and weekly returns were investigated for normality using all the preceding procedures of exploratory data analysis and statistical tests used on the monthly data.Following are some of the graphs produces for Citigroup so as to give an example of the results found across all the assets returns. A number of tables are also supplied:

Looking at the kurtosis values for the log return data at weekly and daily intervals the Wal-Mart kurtosis figures really stand out. At a monthly interval it was the largest of all the assets but as the frequency increases it completely dwarfs the other kurtosis figures. This together with the weekly and daily skewness both being near zero tells us that the Wal-Mart log returns distribution shows a symmetrical, peaked and very kurtotic distribution making it almost certainly non-Normal.

At a more frequent interval it can been seen that the returns are differently distributed to the normal. In particular from the graphs and tests the returns show definite excess kurtosis proving the idea that they are heavy tailed. There is also noticable negative skewness which indicates that losses are more extreme than gains.

So far we have seen strong evidence towards the weekly and in particular the daily returns being skewed to the left and having considerable excess

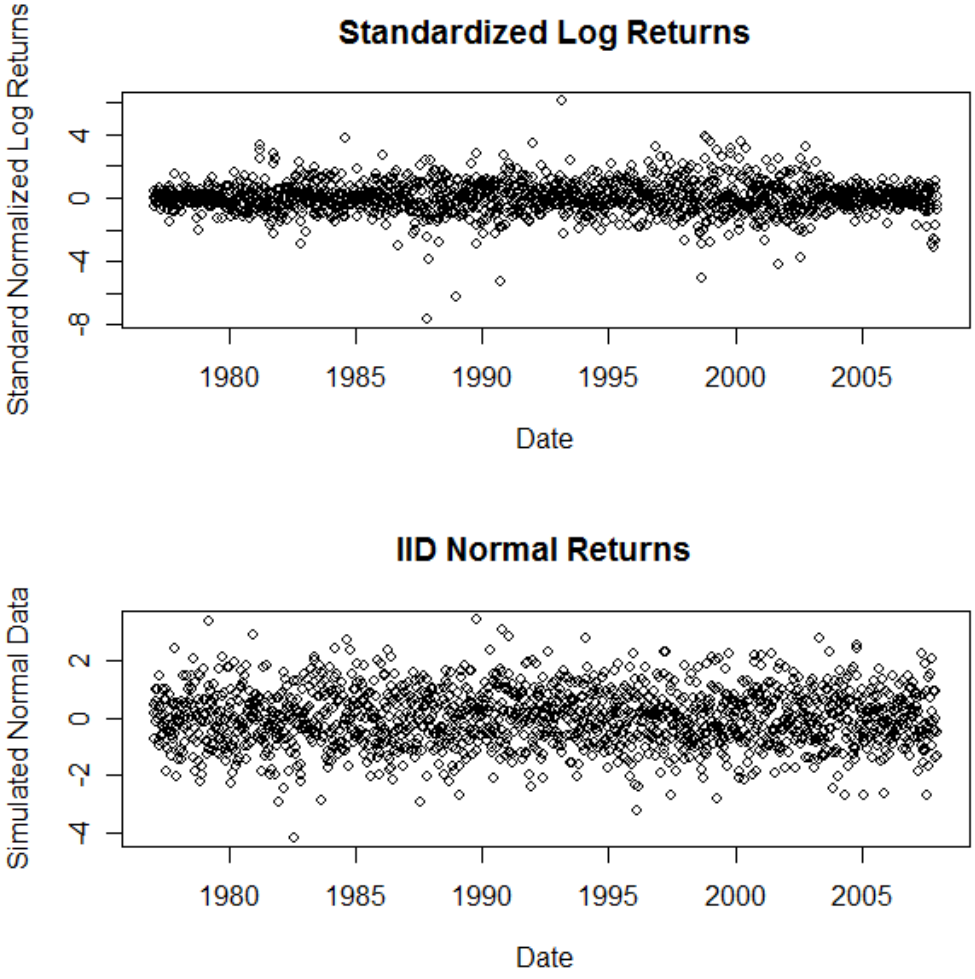

Figure 3.24: Plot of Citigroup Weekly Std Normal Log Returns v Simulated Normal Data: Note the difference in the scales of the Y-axes

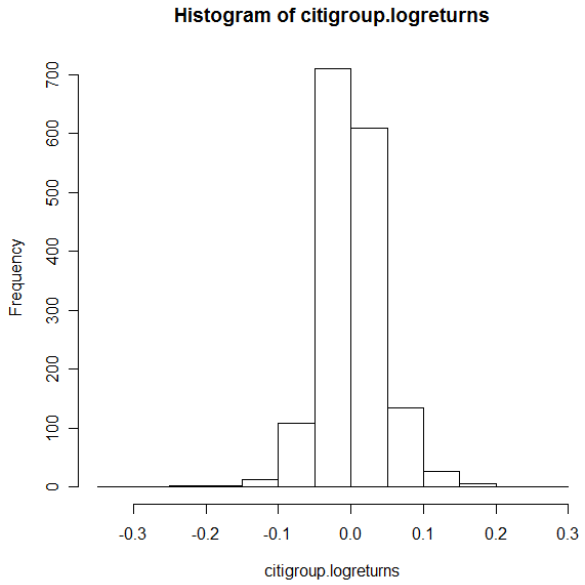

Figure 3.25: Histogram of Citigroup Weekly Log Returns

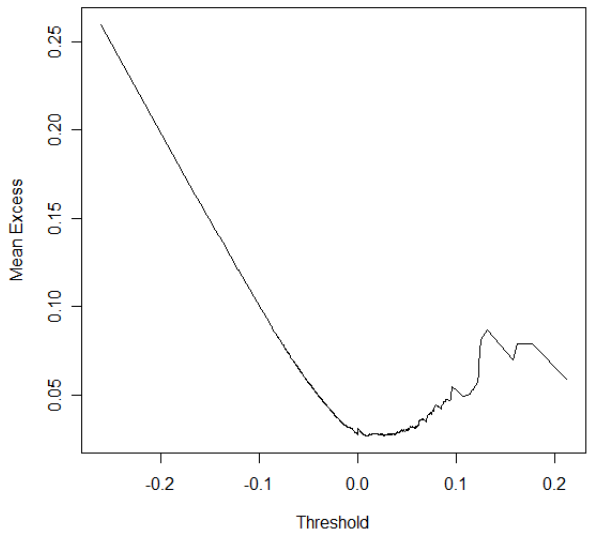

Figure 3.26: Mean Excess plot of citigroup Weekly Losses

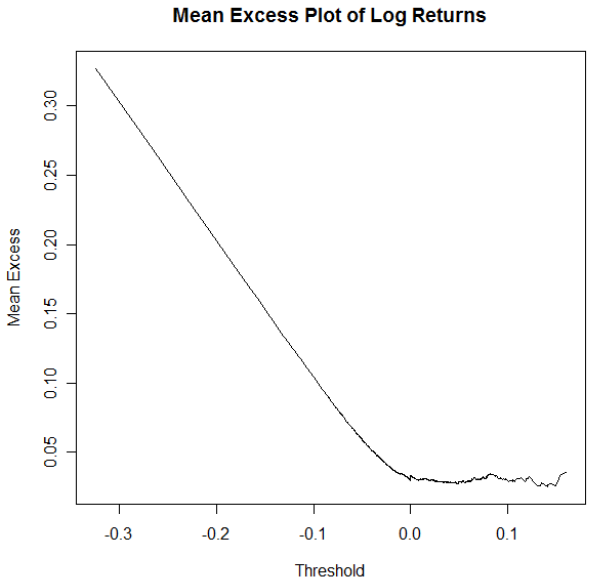

Figure 3.27: Mean Excess Plot of Citigroup Weekly Log Returns

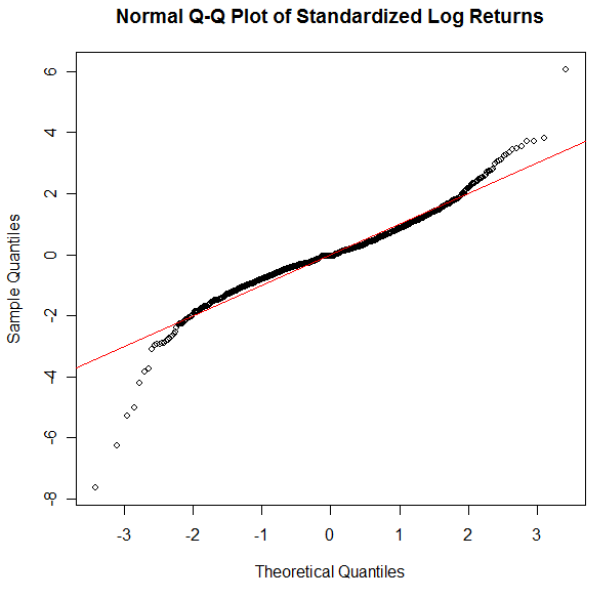

Figure 3.28: QQ plot of Citigroup Weekly Std Log Returns

Company	Minimum	Maximum	Mean	Median
Boeing	-0.40596	0.223144	0.002423	0
Citigroup	-0.324723	0.261645	0.002124	0
General motors	-0.244433	0.171598	0.001055	0
Intel	-0.451985	0.22871	0.003838	0.005102
Wal Mart	-0.693147	0.693147	0.003719	0
Dow Jones	-0.168969	0.167296	0.00097	0.002551

Figure 3.29: Table of Statistics for Weekly Data(1)

Company	Variance	Stdev	Skewness	Excess Kurtosis
Boeing	0.002142	0.046284	-0.269017	3.886343
Citigroup	0.001827	0.042746	-0.268218	5.557949
General motors	0.001393	0.037317	-0.110492	2.664814
Intel	0.003303	0.057475	-0.816253	4.882574
Wal Mart	0.003677	0.060635	-0.046496	32.014657
Dow Jones	0.000599	0.024466	-0.492163	5.70685

Figure 3.30: Table of Statistics for Weekly Data(2)

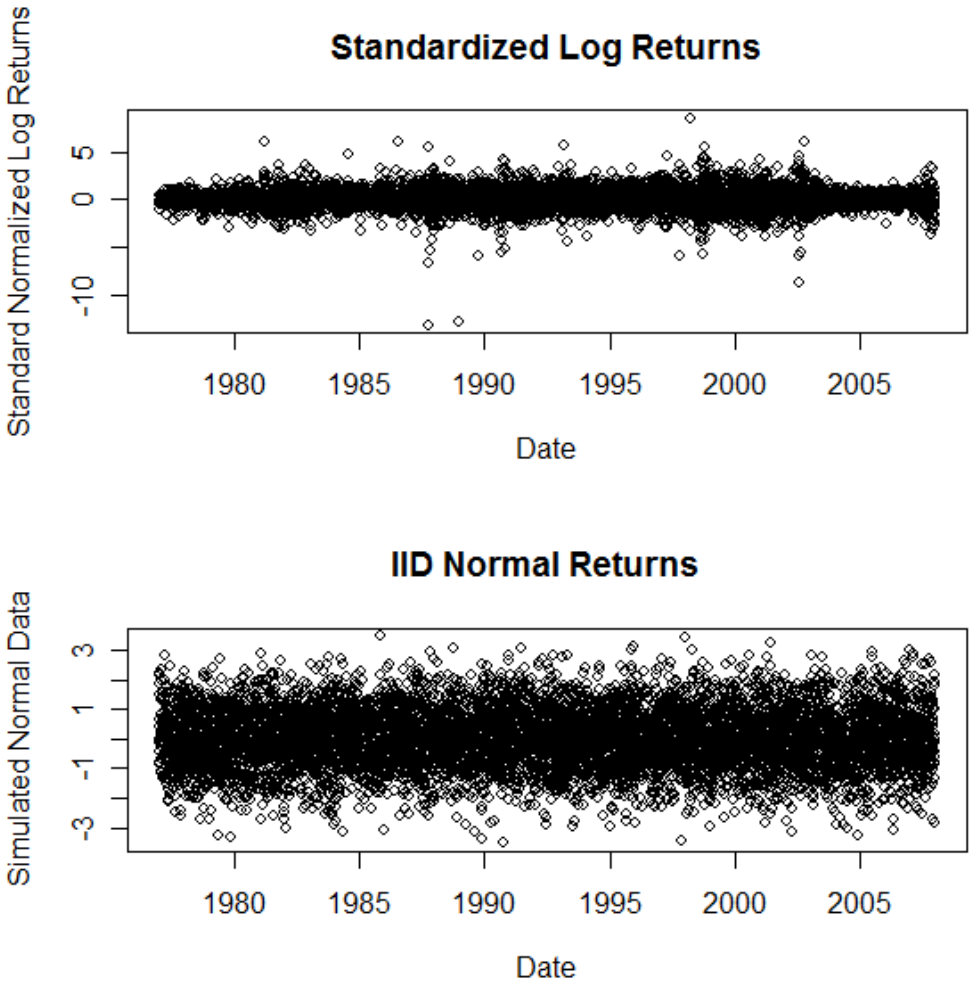

Figure 3.31: Plot of Citigroup Daily Std Normal Log Returns v Simulated Normal Data: Note the difference in the scales of the Y-axes

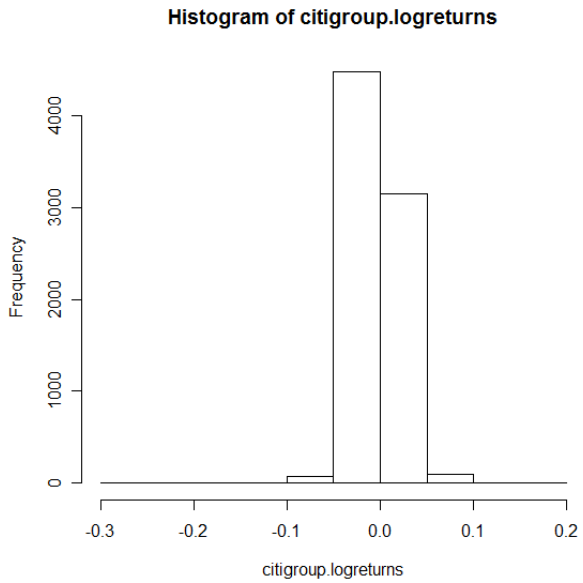

Figure 3.32: Histogram of Citigroup Daily Log Returns

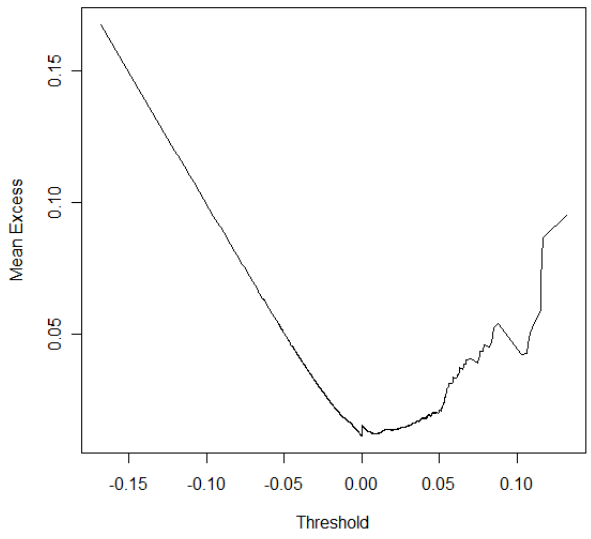

Figure 3.33: Mean Excess plot of citigroup Daily Losses

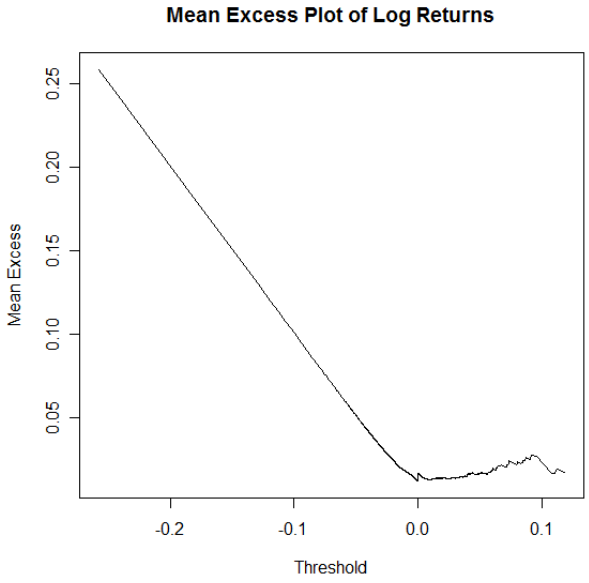

Figure 3.34: Mean Excess Plot of Citigroup Daily Log Returns

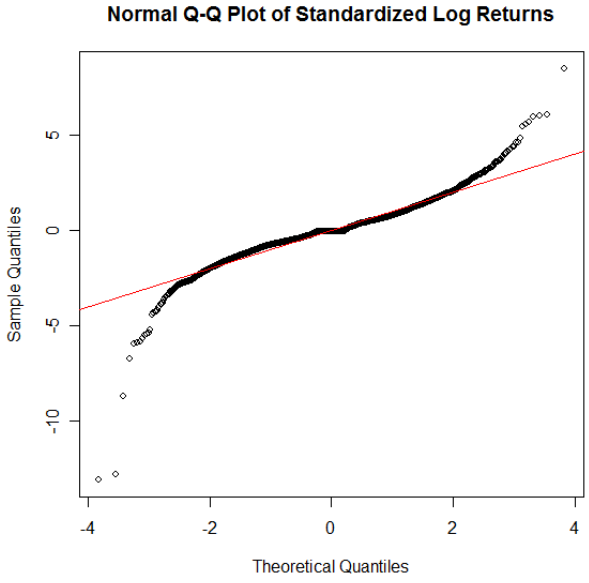

Figure 3.35: QQ plot of Citigroup Daily Std Log Returns

Company	Minimum	Maximum	Mean	Median
Boeing	-0.40596	0.169466	0.001868	0.00285
Citigroup	-0.223144	0.261645	0.003106	0.00243
General motors	-0.244433	0.171598	0.000374	0
Intel	-0.294075	0.22871	0.003483	0.004126
Wal Mart	-0.160468	0.167764	0.002386	0.001811
Dow Jones	-0.153881	0.080898	0.001669	0.002981

Figure 3.36: Table of Statistics for Daily Data(1)

Company	Variance	Stdev	Skewness	Excess Kurtosis
Boeing	0.00178	0.042194	-1.216008	11.355649
Citigroup	0.001986	0.044567	0.109901	3.271056
General motors	0.002121	0.046051	-0.137726	1.963576
Intel	0.003049	0.055215	-0.521129	2.360537
Wal Mart	0.001483	0.038512	0.14558	1.411223
Dow Jones	0.000438	0.020925	-0.614823	3.962129

Figure 3.37: Table of Statistics for Daily Data(2)

kurtosis. The indication is that returns have a greater chance of realising extremes, especially negative extremes, i.e. losses, than the normal distribution can predict. We are drawn to the conjecture that the returns are not sampled from a Gaussian distribution. So then if they are not Gaussian what distribution do they follow or is it possible that they are distributed as a combination of distributions?

As seen from the normal Q-Q plots the centre of the data displayed normality while it was in the tails that we viewed deviation from the norm. It is therefore the extreme values in the tails that we will be most interested in. We shall examine this in greater detail in the next section on extreme value theory which is concerned only with data outside the centre. We shall use this to gain an insight into the behaviour of the tails in particular the negative tails containing the losses.

Chapter 4

Extreme Value Theory Approach

The previous section has shown the log returns demonstrate heavy tails. A method of capturing the behaviour in the tails is applied through the extreme value theory approach.

4.1 Extreme Value Theory

Extreme Value Theory is concerned with the study of the asymptotic behaviour of extreme observations of a random variable. In more conventional techniques due to the contribution of the tails being relatively smaller than that of the observations in the centre of the distribution the tails were neglected. Extreme Value Theory (EVT) takes a contradictory approach emphasising the importance of the tails distributions. Consequentially measure-

ments concerning extremes, such as the quantiles required in Value-at-Risk (VaR), can be estimated more accurately using EVT based techniques than more conventional approaches.

One of the benefits of EVT is that it does not require that a priori assumptions be made about the underlying distribution from which the empirical data was sampled. Due to the Fisher-Tippett Theorem (1928), or the extremal types theorem as it is also known, possible suitable classes of distribution can be identified under EVT to model the actual underlying return distribution. This permits figure estimation processes, such as the previously mentioned VaR, to be carried out without first making a priori assumption concerning the return distribution.

The classical Extreme Value theory (EVT) is used in the study of the asymptotic behaviour of extreme observations (maxima or minima of n random realisations).

Let X be a random variable with the density f and the cumulative distribution function (cdf) F. If $X_1, X_2, \ldots, X_n$ are a set of n independent realisations of this random variable then the extreme observations are defined as follows:

$$\begin{aligned} Y_n &= max\{X_1, X_2, ...X_n\} \\ Z_n &= min\{X_1, X_2, ...X_n\} \end{aligned}$$

eqnarray* EVT looks at the distributional traits of Y_n and Z_n as n grows. The exact distribution of the extreme values is degenerate as n tends to infinity. To get a distribution from Y_n and Z_n that is non-degenerate they

are standardized with the use of a location parameter a an and a positive scale parameter b. The distribution of these standardized extrema

$$\frac{Y_n - a_n}{b_n} \qquad \frac{Z_n - a_n}{b_n}$$

is non-degenerate in the limit. It is worth noting that the maximum and minimum are related by the following equality

$$min\{X_1, X_2, ...X_n\} = -max\{-X_1, -X_2, ... - X_n\}.$$

This means that we can focus on the properties of the maxima as any conclusions will be analogous to the minima.

4.1.1 Fisher-Tippett Theorem

Also known as the Extremal Type Theorem. If there exist the normalizing constants $a_n > 0$ and $b_n \in R$ such that

$$\frac{Y_n - a_n}{b_n} \longrightarrow H \qquad asn- > \infty$$

for some non-degenerate distribution H, then H must be of one of 3 possible extreme value distributions. The 3 types of distribution are

1. The Gumbel or Type I Distribution

2. The Fréchet or Type II Distribution

3. The Weibull or Type III Distribution

For a more particular analysis of these three types consult Embrechts et al (1997).

The first proof of this was by Gnedenko (1943) and so Fisher and Tippett sometimes share the theorem title with him. A simpler proof was given by De Haan (1970), although an even more simplified version was later proposed by Weissman (1977).

The Gumbel set of distributions are used to describe thin tailed distributions and include those such as the Gaussian and the log normal distributions. The Fréchet distributions include the stable Pareto and the students-t distributions and are used to describe heavy tails. The Weibull is used when a distribution has a finite endpoint and no tail. Weibull distributions include the uniform and beta distributions.

A random variable X, and its underlying distribution F are said to belong to the maximum domain of attraction of extreme value distributions denoted by $X \in DA(H)$ where H represents extreme value distributions. This terminology can also be used to summarize the types of distributions as shown

$$\begin{aligned} Normal, lognormal, exponential &\in DA(Gumbel) \\ Pareto, students-t, Cauchy &\in DA(Fréchet) \\ Uniform, Beta &\in DA(Weibull) \end{aligned}$$

4.1.2 Generalisalized Extreme Value Distribution

The 3 forms of extreme value distribution can be encompassed into a single parametric representation shown as shown by Jenkinson and Von Mires (1955). This representation is named the Generalized Extreme Value (GEV) distribution defined as follows

$$H_\xi(x) = exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right)$$

where $1 + \xi x > 0$.

Therefore

$$\begin{aligned} & for \quad \xi > 0, \qquad x > -\frac{1}{\xi} \qquad i.e. Fréchet \\ & for \quad \xi < 0, \qquad x < \frac{1}{\xi} \qquad i.e. Weibull \\ and \quad & for \quad \xi = 0, \qquad x \in R \qquad i.e. Gumbell \end{aligned}$$

ξ is called the tail index and controls the distributions tails.

It can be seen that all the common, continuous distributions important in statistics belong to the domain of attraction of the GEV distribution. This demonstrates the generality that the Fisher-Tippett theorem allows.

4.1.3 General Pareto Distribution

The generalised Pareto distribution is defined as:

$$G_{\xi,\beta(u)}(x) = \begin{cases} 1 - (1 + \frac{\xi x}{\beta})^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - exp\left(\frac{-x}{\beta}\right) & \text{if } \xi = 0 \end{cases}.$$

and the support of x is $x \geq \frac{-\beta}{\xi}$ if $\xi \geq 0$, and $0 \leq x \leq \frac{-\beta}{\xi}$ when $\xi < 0$.

4.2 Peak Over Threshold Method

4.2.1 Introduction to Peak Over Threshold

In the study of heavy negative tailed distributions which are encountered during empirical examination of asset log returns EVT is a basis for the so-called Peak Over Threshold Method (POT). The POT method is commonly used when asymmetry is observed in the distribution of the data as observed with financial assets log returns. When using POT it is expected that the centre of the data follows a Gaussian distribution while the tails are significantly different from the Gaussian. One uses POT to get an accurate estimate of the right tail of the distribution of potential losses which correspond to the left tail of the log returns. The upper tail of the losses is expected to exhibit significant tail thickness which can be speculated at from the asymmetry of the underlying distribution.

It is reasonable here to introduce losses as negative log returns. Simply:

$$losses = -(logreturns)$$

The only affect this has on the return basic statistics is to simply change the sign of the mean, the sum and the skewness, while taking the opposite signs and reversing the maxima and minima and the first and third quantiles. Importantly with regards the assumption of no normality all other features such as the standard deviation and the kurtosis remain unchanged.

The POT method is then applied to get an accurate estimate of the right tail of the losses. Note that the right tail of log returns is also heavy tailed

but not of interest for risk managers since this corresponds to profits. It is also less skewed and therefore less extreme than the tail of the losses.

The POT method attempts to estimate the tails of the underlying distribution. A particular threshold is identified that is used to define the starting point of the tail of the distribution. An estimate of the distribution of the excesses over the particular threshold is obtained. There are 2 common methods for estimating the distribution. The first is a semi-parametric model based on the Hill estimator as described by Danielsson and de Vries (1997). The trouble with this method is that it requires the assumption of a fat tailed underlying distribution. The second approach is the fully parametric model based on the general Pareto distribution (McNeil and Frey, 1999). As this method does not require any assumption about the underlying tails it can be applied to any distribution. It makes use of the Pickand, Balkema and de Haan theorem (which is discussed below) to fit a GPD to the tail which has been defined to be the excesses over a particular threshold. It is this second and more easily applicable method that will be used to estimate the tail of the return losses.

4.2.2 Pickands-Balkema-De Hann Theorem

Letting X be a random variable with distribution F, if $X_1, X_2, \ldots X_n$ are a set of n independent realisations of this random variable then the distribution function of excess over a certain threshold u is defined by

$$F_u(x) = P\{X - u \leq x | X \geq u\} = \frac{F(x+u) - F(u)}{1 - F(u)}$$

The Pickand, Balkema and de Haan theorem says that if the distribution F which is a domain of attraction of $H_\xi(x)$ then there exists a positive measurable function of the threshold $\beta(u)$ such that

$$\lim_{u \to k} sup_{0 \leq x \leq k-u} \left| F_u(x) - G_{\xi,\beta(u)}(x) \right| = 0$$

where $G_{\xi,\beta(u)}(x)$ denotes the generalised Pareto distribution (GPD).The theorem says that the distribution of the excesses over the threshold tend to the GPD as the threshold u becomes large. This statement is based on the assumption that the underlying F belongs to the domain of attraction of the GEV distribution.

4.2.3 POT Using GPD Approach

Peak Over Threshold provides a structure for estimating the tails, in our case the losses, of a distribution without making a priori assumptions about the tail thickness. A threshold is chosen as defining the start point of the tail and the POT method then estimates the distribution of the excesses beyond the threshold. The distribution of excesses over a sufficiently high threshold u on the underlying return distribution F is defined by

$$F_u(y) = Pr\{X - u = y | X > u\}$$

A sufficiently high threshold provides an optimal balance between the bias of the model which is increased as the threshold becomes lower and the variance of it which grows as the threshold does due to the lack of data points (McNeil,

1997). The above equation can be rewritten in terms of the underlying F as

$$F_u(y) = \frac{F(y+u) - F(u)}{1 - F(u)}$$

According to Pickands, Balkema and de Haan the excess distribution can be approximated well by a general Pareto distribution as the threshold becomes large:

$$F_u(y) \rightarrow G_{\xi,\beta(u)}(y) \quad as \quad u \rightarrow k$$

Setting $x = u + y$ and using the above two statements the distribution of the excess function can then be restated as below providing u is sufficiently high:

$$\begin{aligned} G_{\xi,\beta(u)}(y) &= \frac{F(x) - F(u)}{1 - F(u)} \\ F(x) &= (1 - F(u))G_{\xi,\beta(u)}(y) + F(u) \quad for \quad x > u \end{aligned}$$

Historical simulation is used to get the empirical estimate of $F(u)$

$$\hat{F}(u) = \frac{n - N_u}{n}$$

where N_u is the number of log returns exceeding the threshold u.(McNeil, 1999). Using this estimate for $F(u)$ and maximum likelihood estimates to obtain the GPD parameters ξ and β allows the following tail estimate formula to be achieved:

$$\hat{F(x)} = 1 - \frac{N_u}{n}\left(1 + \hat{\xi}\left(\frac{x-u}{\hat{\beta}}\right)^{\frac{-1}{\hat{\xi}}}\right)$$

The tail estimate formula can be seen as an augmented form of historical simulation using EVT. It must also be recognised that this formula is build on the belief that the data is identically distributed. Although it works best

on independent data the tail estimate formula can also be applied to weakly dependent data with satisfactory effect (McNeil, 1999).

A useful result of this formula to estimate the tail is that it can be inverted to give a tail quantile function:

$$\hat{q}_p = u - \frac{\hat{\beta}}{\hat{\xi}} \left(\frac{n}{N_u} (1-p)^{-\hat{\xi}} - 1 \right)$$

The tail quantile function can be used to get that p_{th} quantile where p is a given probability.

4.2.4 Application of POT to the Tails

The Pickand, Balkema and de Haan theorem suggests GPD as a natural choice when trying to model the distribution of excess over a sufficiently high threshold. As previously mentioned the choice of this threshold is a trade off between bias and variance. A threshold that is too low will include data values from the centre of the return distribution which will influence and cause bias in the model making it invalid for the tail. Too high of a threshold will result in not enough data points to estimate the parameters properly using GPD. Due to insufficient data the influence of a few outliers will cause too much variance to make the model meaningful.

There is no definite correct technique for choosing a sufficient threshold (Sarma, 2002). Analysts take many varied approaches to this task and it is most of the time a question of personal choice. Gavin (2000) used an arbitrary 90% confidence interval taking the largest 10% of returns to be extreme observations. Neftci (2000) chose a threshold of 1.65 times the unconditional

variance of the residuals to mark the start of the tail. McNeil and Frey (1999) used the same mean excess plot described earlier as their method to select their threshold. Although all three methods have held up well in their respective empirical studies the techniques we will use are the Threshold Choice plot and the Mean Residual Life plot, invoked by tcplot(data) and mrlplot(data) respectively to aid picking the optimal threshold.

The Threshold Choice (TC) plot uses maximum likelihood estimation to get estimates for the shape and modified scale parameters which are then plotted against their corresponding thresholds. The scale is modified as it is obtained by subtracting the shape multiplied by the threshold. If a threshold is appropriate to be used in the POT method then the parameter estimates are approximately constant above it. In our graphs there is also displayed a 95% confidence interval indicated by the whiskers either side of the plotted point. The Mean Residual Life (MRL) plot graphs the set of points

$$\{u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x(i) - u)\}$$

where $x(i)$ is an observation that exceeds the threshold. It looks at the average residual between $x(i)$ and the threshold and it should be approximately linear for $u > u_0$ if the $x(i)$ that exceed u_0 are from a generalized Pareto distribution. Again a 95% confidence interval is shown based on the assumed approximate normality of the sample means.

Looking at the plots shown for the monthly log returns of Boeing the TC plots of the parameter estimates becomes approximately linear between 0 and 0.1. It should also be notes that the confidence intervals begin to get

wide after this point due to a lack of data points. This seems like a reasonable region to pick a threshold to start the tail. For Citigroup, General Motors, Intel and Wal-Mart we take the same approach looking for linearity in the MRL plot and constant threshold values in the TC plot above a certain threshold. It should be noted that while theoretically to follow a GPD the plots should display the above characteristics, in reality these traits are much less distinct with the graphs showing erratic behavior at times.

After obtaining various estimates for the starting points of the tail it is sensible to plot the fitted GPDs for a threshold range around this region. As an example for Boeing starting with the threshold 0.01 and incrementing this by 0.01 each time we fit the data above the threshold to a GPD and plot this. The result is four graphs that indicate the fit of the data to a GPD where the parameters have been fitted by the MLE procedure. The three graphs we will have most interest in are the Probability-Probability (P-P) plot, the Quantile-Quantile (Q-Q) plot and the Density plot. All the graphs are fitted with a 95% confidence interval also and they allow us to see data above which threshold adheres best to a GPD. The P-P and Q-Q plots also have a positive sloped diagonal line running through them to which the data should reside close to if it is a good fit to the GPD under the maximum likelihood estimates.

The optimal threshold choice, as well as satisfying the TC and MRL plots, should have P-P and Q-Q plots that are reasonably linear and a density plot that is smooth and follows the data to an acceptable extent. When the optimal threshold is chosen a final GPD fitting occurs and the estimated

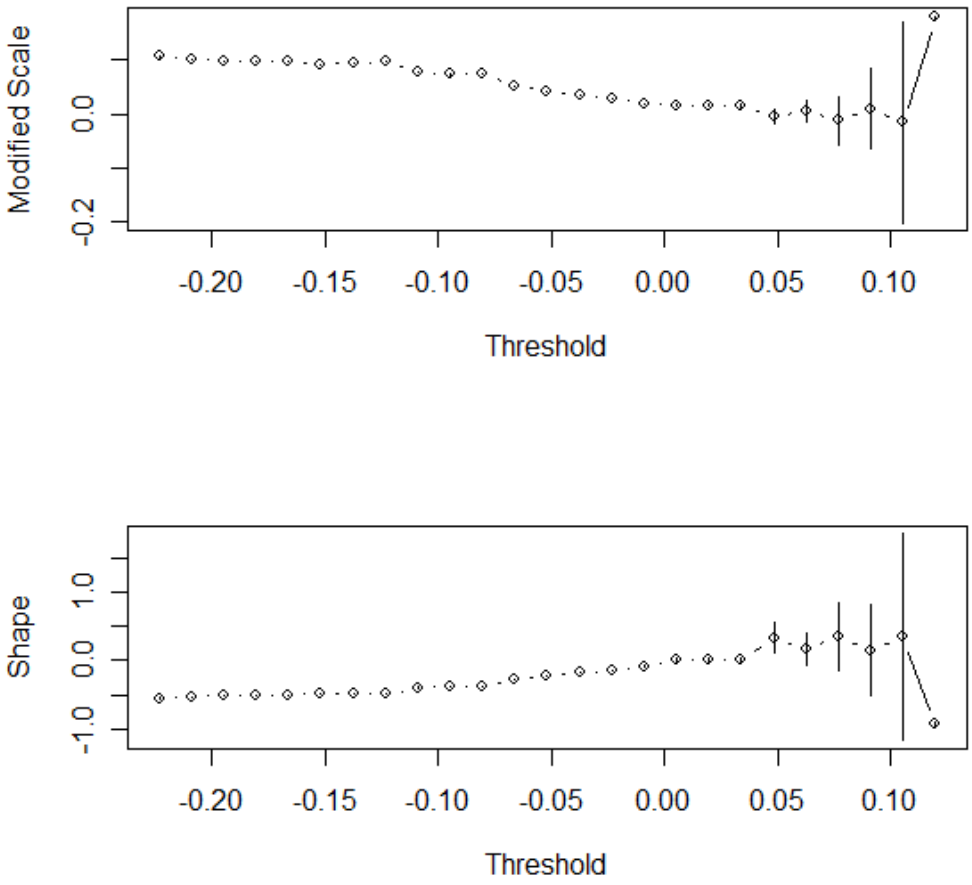

Figure 4.1: Threshold plot of Boeing Daily Returns

scale and shape parameters are defined. The table of chosen thresholds and related scale and shape parameters is shown below.

The table and graphs show that the tails can be appropriately fitted to a GPD with estimated scale and shape parameters as given. The threshold

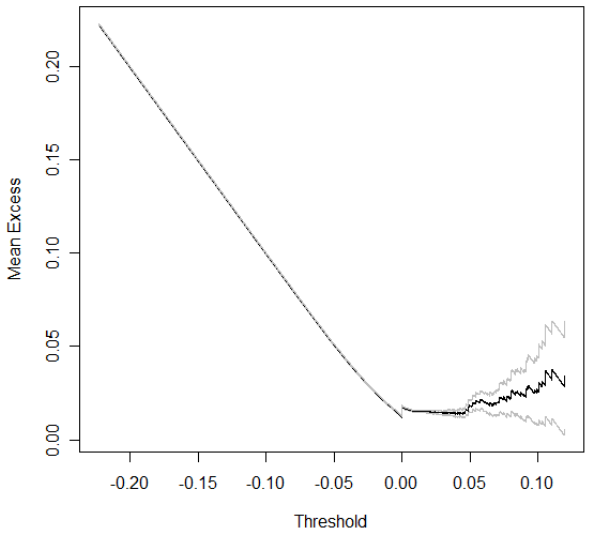

Figure 4.2: Mean Residual Loss plot of Boeing Daily Returns

that were chosen have are above the 95_{th} percentile of the data. This means that only 5% or less of the data is considered to be an extreme loss residing in a tail of the data that can be modelled as GPD. As discussed by McNeil (1999) this should be an acceptable threshold level. As can be seen the P-P and Q-Q plots can be seen to be linear for the most part while a few show specific deviations from the line at the upper end of the Q-Q plots.

Of concern are the GPD plots of the Wal-Mart losses positive tail. They show very strange patternsof groups of horizontal lines across the diagonal Q-Q plot. At a monthly level this was less evident but at a daily frequency the patterns are very strong. From the plot of the daily log returns we can see some very unusual activity for a number of years after the stock was first

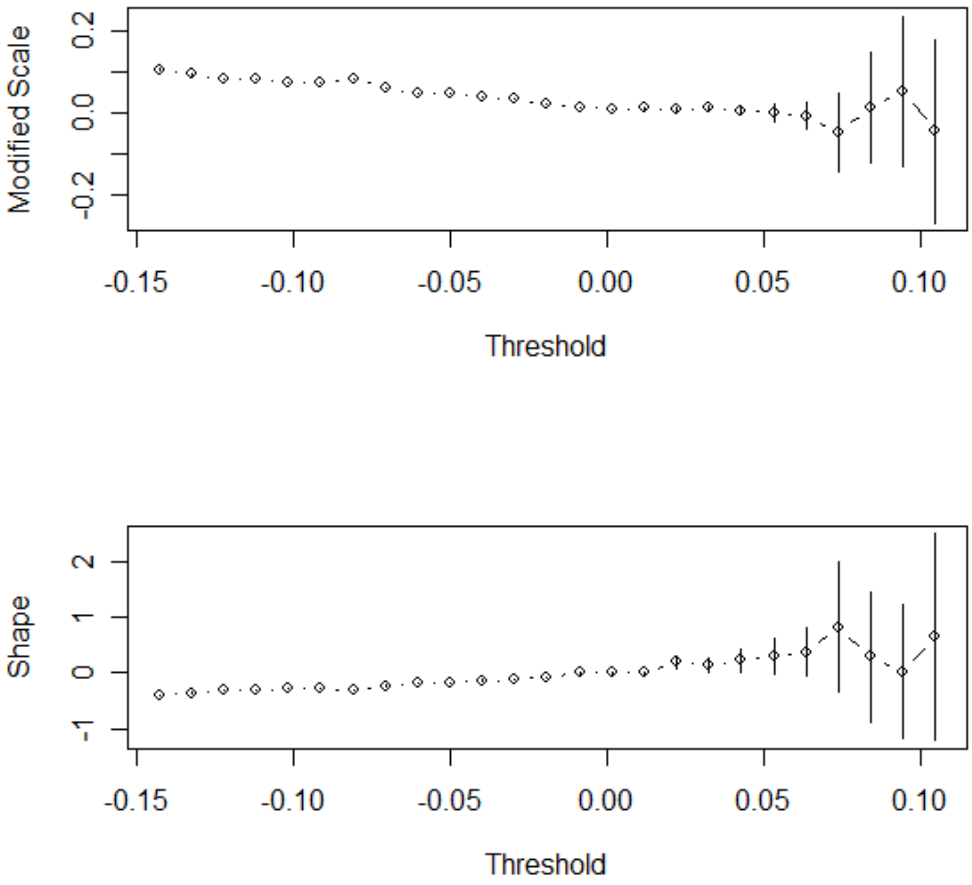

Figure 4.3: Threshold plot of Boeing Daily Returns

offered on the stock exchange.

A look back at the raw data shows that the Wal-Mart stock was split quite substantial giving the adjusted close price a value of only a few cents. Due to this any small change in this value of even a cent makes the log returns

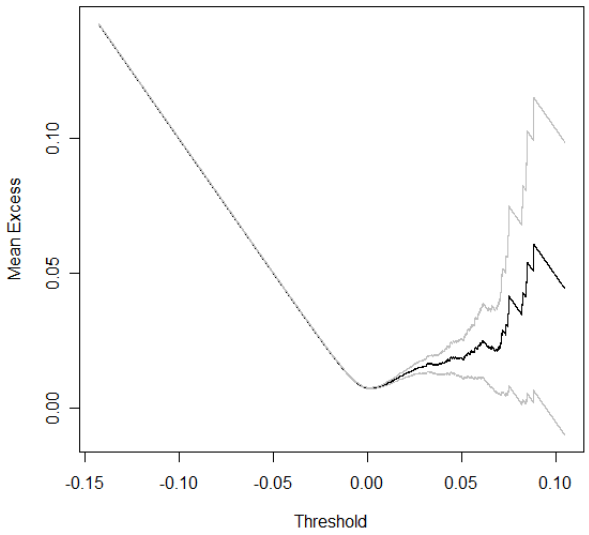

Figure 4.4: Mean Residual Loss plot of Boeing Daily Returns

look very large. From the plot of the log returns it is seen that not until the early part of the 1980's that the adjusted stock price becomes reasonable enough to deal with. A Q-Q plot of the data over 1st Jan 1990 to 31st Dec 2007 shows behaviour similar to that of the other stocks therefore the odd behaviour in the GPD over the full time period can be accounted for as a peculiarity in the stock due to a 100% stock split.

We will then expect that a the set of data outside the period up until the early 1980's will display the same characteristics as the other stocks over their whole time period. If for instance we take the negative tail of the Wal-Mart returns over the 1990 to 2007 period already used we expect that a GPD will model this well. The following graph in figure!!! demonstrates this.

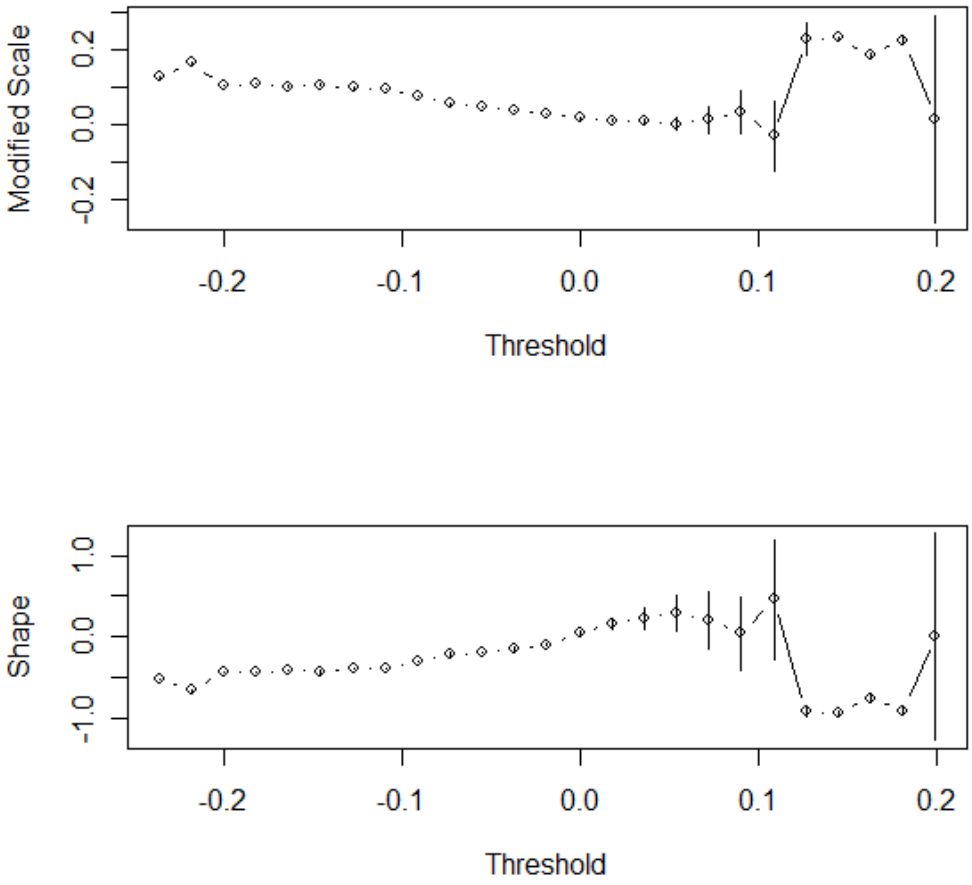

Figure 4.5: Threshold plot of Boeing Daily Returns

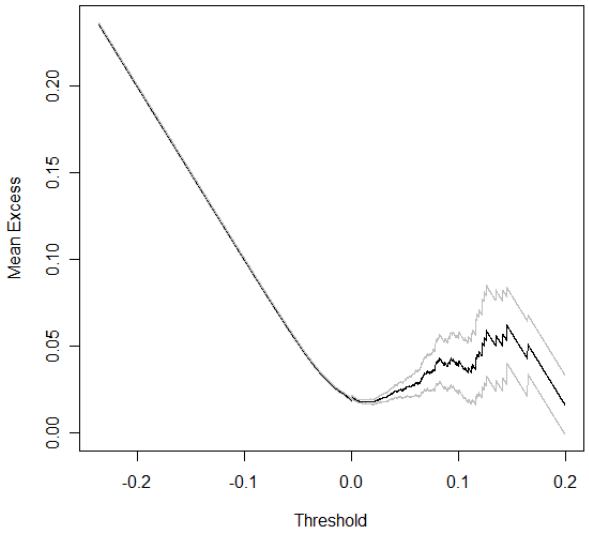

Figure 4.6: Mean Residual Loss plot of Boeing Daily Returns

Daily	scale	shape	std.err.scale	std.err.shape	threshold	% over threshold
Boeing	0.011241	0.1979966	0.0008281	0.056945849	0.04	0.0356711
Citigroup	0.011739	0.2826663	0.001119497	0.07508329	0.035	0.0315776
General Motors	0.01043	0.1722519	0.000745527	0.05328208	0.03	0.03463465
Intel	0.019434	0.2140187	0.002029016	0.08052047	0.045	0.03949797
Walmart	0.015153	0.8656469	0.001385158	0.08885873	0.03	0.0483238
Dow Jones	0.009995	0.2140334	0.000617775	0.04953602	0.02	0.03135994

Figure 4.7: Table of Parameters Estimates for Selected Threshold

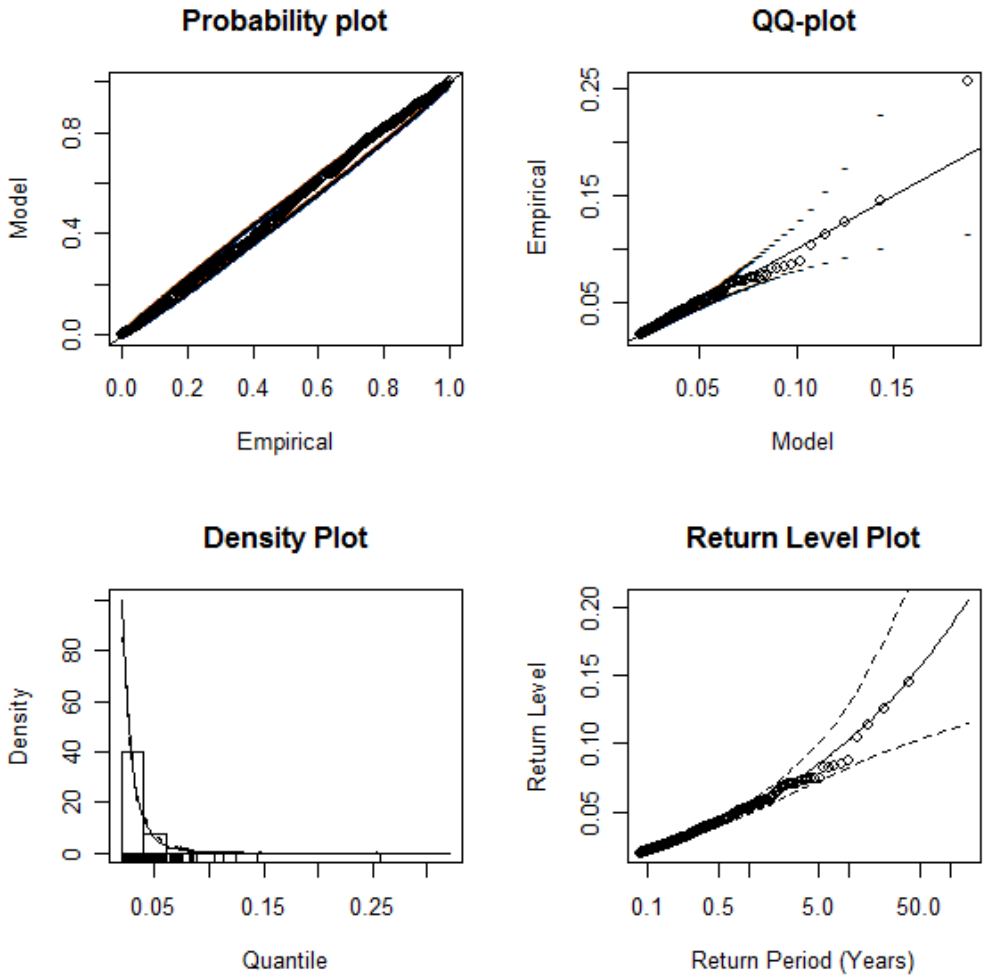

Figure 4.8: Plots of Dow Jones Fitted Excesses

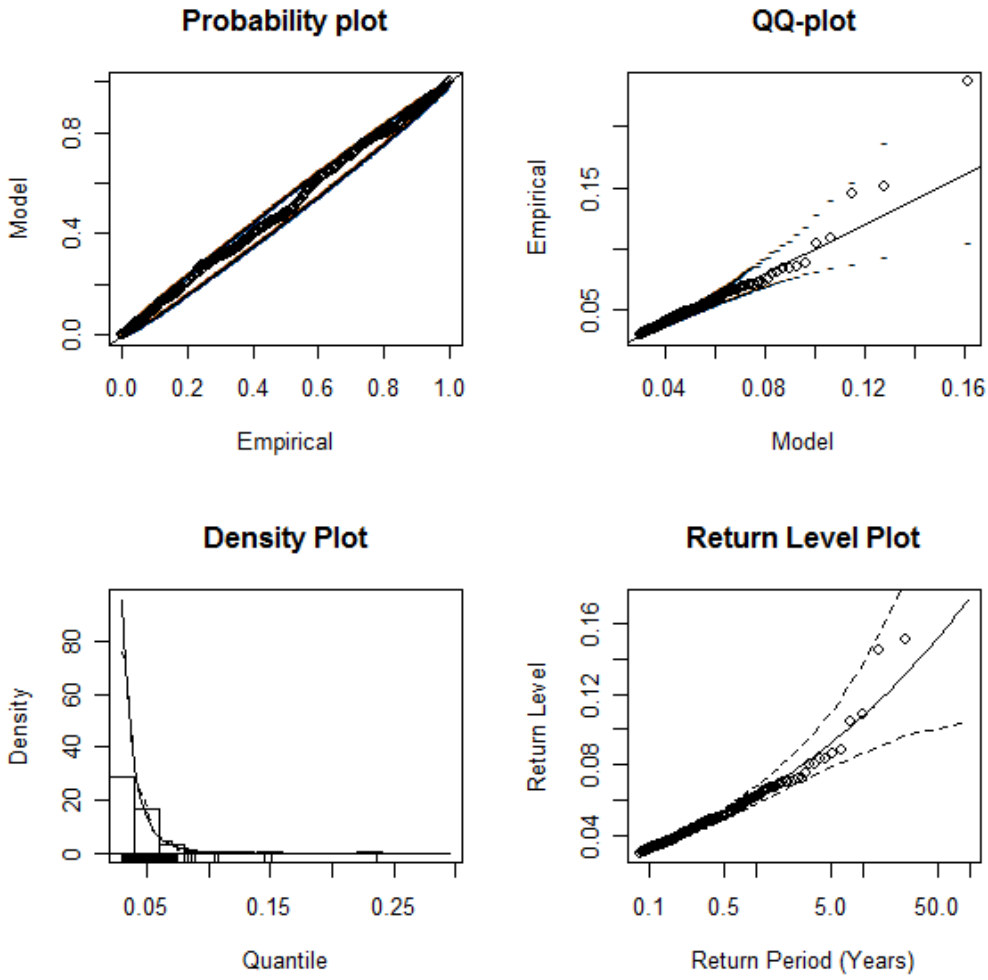

Figure 4.9: Plots of General Motors Fitted Excesses

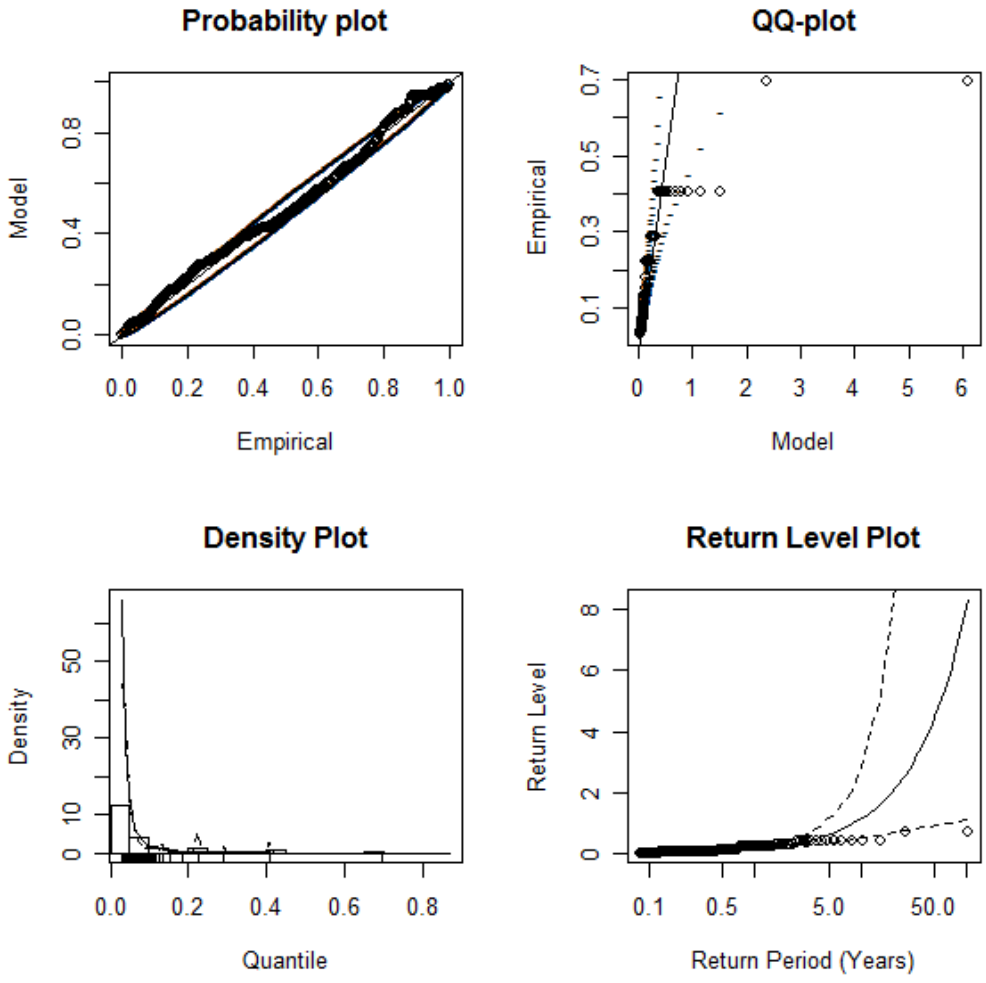

Figure 4.10: Plots of Wal-Mart Fitted Excesses

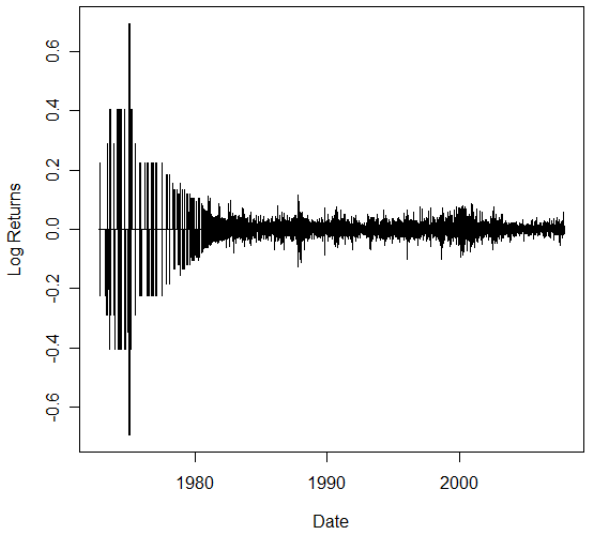

Figure 4.11: Graph of Wal-Mart Log Returns

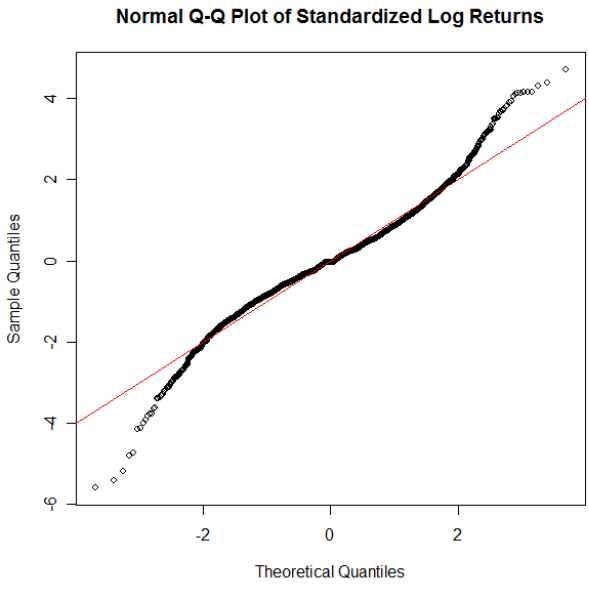

Figure 4.12: Q-Q Plot of Wal-Mart Log Returns 1990-2007

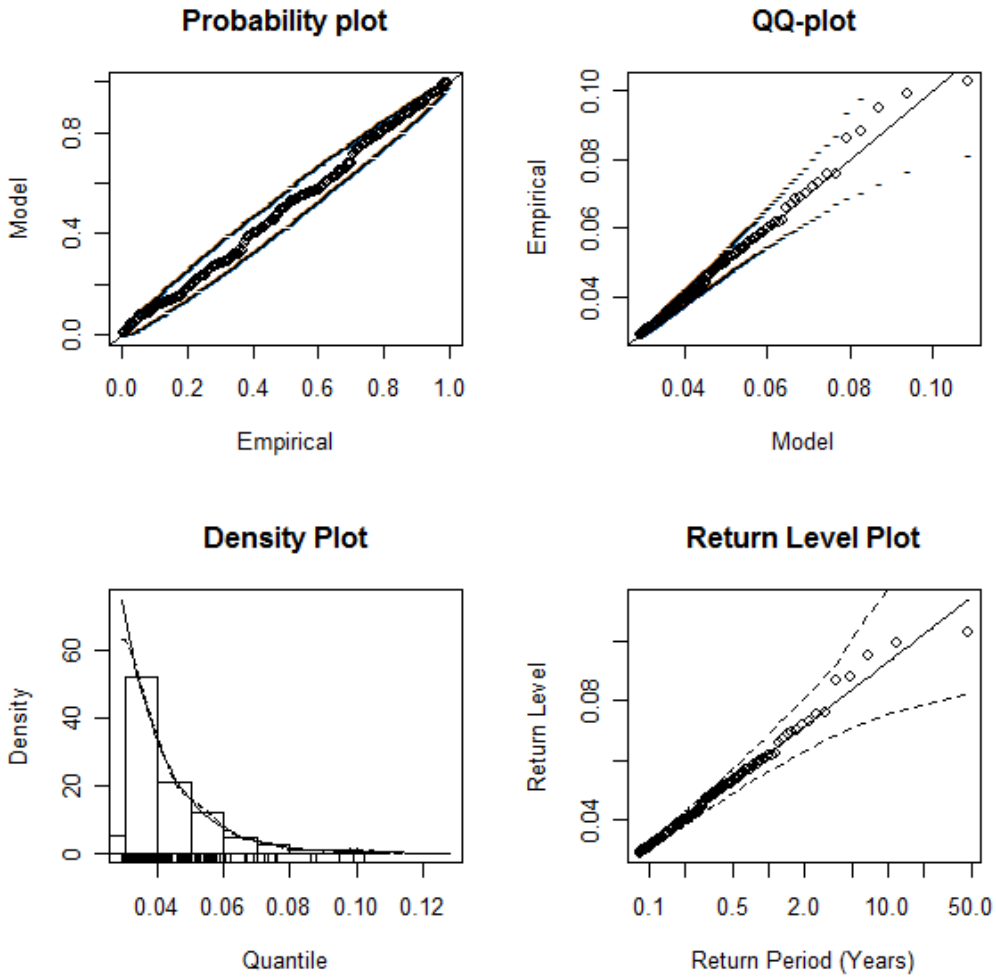

Figure 4.13: Plots of Wal-Mart Fitted Excesses 1990-2007

Chapter 5

Time Series Approach

It has been seen that the POT method of fitting a GPD to the tail works considerably well. We stated earlier that this tail estimate is best applied under the assumption of independence but is also suitable under weak dependence. We have noticed though that the log returns show fluctuating periods of high and low volatility termed volatility clustering. This volatility clustering indicates that in fact there is dependence present in the return series which leads to the opinion that the previously described random walk approach may not be fully applicable to it.

Another approach is that of time series analysis. A time series is an ordered sequence of values of a variable at equally spaced time intervals. It can take into account the possible internal structure of the series in particular auto- or serial-correlation and periodic variation in the data. Therefore it seems a suitable approach when investigating a dependent series. In our case we will be dealing with univariate time series'. In our approach to use time

series analysis we will discuss the following aspects:

- stationarity
- correlation in particular autocorrelation
- white noise in a linear time series

5.1 Stationarity

A time series is strictly stationary if the joint distribution of $(r_{t_1}, r_{t_2}, ..., r_{t_k})$ is identical to $(r_{t_1+t}, r_{t_2+t}, ..., r_{t_k+t})$ for all values of t where the joint distribution is defined by: $\mathrm{F}_{(X,Y,...)}(x, y, ...; \theta) = Pr(X \leq x, Y \leq y, ...)$

This means that the joint distribution is invariant with respect to time changes. This is basically impossible to verify through empirical study so the more commonly used assumption of weak stationarity is accepted.

Under weak stationarity a time series has time invariant first and second moments. That is to say that the mean and lag-ℓ covariance are invariant under time shift. In particular the mean is constant with the lag-ℓ covariance only depending on ℓ where ℓ is an integer. This statement can be given by:

$$\begin{aligned} E[r_t] &= \mu \qquad \forall t \\ Cov(r_t, r_{t-\ell}) = E[(r_t - \mu)(r_{t-\ell} - \mu)] &= \gamma_\ell \qquad \forall t \quad and \quad \ell \end{aligned}$$

Essentially this means that a stationary series fluctuates about a constant level with a constant variation.

γ_ℓ is called the lag-ℓ serial or autocovariance of the time series r_t. An autocovariance plot can be obtained simply by plotting γ_ℓ against ℓ. The autocovariance has two properties that it is important to be aware of:

$$\begin{aligned} \gamma_0 &= Var(r_t) \\ \gamma_{-\ell} &= \gamma\ell \end{aligned}$$

both of which can be easily proven (Tsay, 2002).

From the definitions it is evident that if a series is strictly stationary and its first and second moments are finite then the series is also weakly stationary. The converse of this is not generally true but under the assumption of normality weak stationarity is analogous to strict stationarity.

To summarize a series is stationary if all its moments are invariant under time shifts while if only the first and second moments are it is only weakly stationary. It is commonly accepted in finance literature that the return series of an asset is weakly stationary.

5.2 Correlation and Autocorrelation.

The correlation is a measure if the linear dependence between random variables. It is measured with the use of a correlation coefficient $\rho_{x,y}$ where X and Y are the random variables in question. The correlation coefficient is

defined as

$$\rho_{x,y} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

where it is assumed that the variance of both X and Y exist. X and Y are uncorrelated if and only if $\rho_{x,y} = 0$ and they are perfectly correlated if $|\rho_{x,y}| = 0$. Two other properties of the correlation coefficient are:

$$\rho_{x,y} = \rho_{y,x} for \quad -1 \leq \rho_{x,y} \leq 1$$

When dealing with a sample $\{x_t, y_t\}_{t=1}^{T}$, where T is the sample size, the sample correlation coefficient is given by

$$\hat{\rho}_{x,y} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sum_{t=1}^{T}(y_t - \bar{y})^2}}$$

where $\bar{x} = \frac{\sum_{t=1}^{T} x_t}{T}$ and $\bar{x} = \frac{\sum_{t=1}^{T} x_t}{T}$

Under a similar definition to autocovariance the lag-ℓ autocorrelation function of r_t can be given. If the series r_t is weakly stationary the lag-ℓ autocorrelation function (ACF) is given by:

$$\rho_\ell = \frac{cov(r_t, r_{t-\ell})}{\sqrt{var(y_t)var(y_{y-\ell})}} = \frac{\gamma_\ell}{\gamma_0}$$

Under weak stationarity it is a function of only ℓ and it describes the linear dependence between r_t and $r_{t-\ell}$ for ℓ an integer. Properties of the autocorrelation function are an extension of the properties defined for simple correlation:

$$\begin{aligned} \rho_0 &= 1 \\ \rho_\ell &= \rho{-\ell} \\ -1 \leq \rho_\ell &\leq 1 \end{aligned}$$

and a weakly stationary series is only not serially correlated if ρ_ℓ for all ℓ.

For a sample of size T the autocorrelation is given as

$$\hat{\rho}_\ell = \frac{\sum_{t=1}^{T}(r_t - \bar{r})(r_{t-\ell} - \bar{r})}{\sum_{t=1}^{T}(r_t - \bar{r})^2} \quad \text{where} \quad 0 \leq \ell \leq T-1.$$

This sample autocorrelation is asymptotically normal with mean 0 and variance $\frac{1}{T}$ for any fixed $\ell \in N^+$ if r_t is an i.i.d. sequence with $E[r_t^2]$ finite. For finite samples $\hat{\rho}_\ell$ is a biased estimator of ρ_ℓ with a bias of the order $\frac{1}{T}$. Obviously then small samples will have a very large bias as expected.

The sample autocorrelation function (ACF) is formed from all the sample autocorrelation coefficients $\rho_1, \rho_2, \ldots$. The sample ACF plays an important role in linear time series analysis as a linear series can be modelled by its sample ACF. The linear characteristics of the time series can be acquired by time series analysis through its sample ACF. The autocorrelation plot is the plot of the γ_ℓ against ℓ.

A very important property of the ACF is that a weakly stationary series is defined through its mean, variance and ACF.

The ACF of the loss distributions of a number of stocks are shown below.Most of the correlations are very small,this is especially true at monthly and weekly frequncies but at daily they become larger. It can be seen from the Wal-Mart ACF for daily returns that at this frequency we cannot ignore the serial correlations of the returns. There are a number of negative correlations that are quiet substantial and suggest further study might be applicable into methods such as ARMA and ARCH models which allow for this skedasticity.

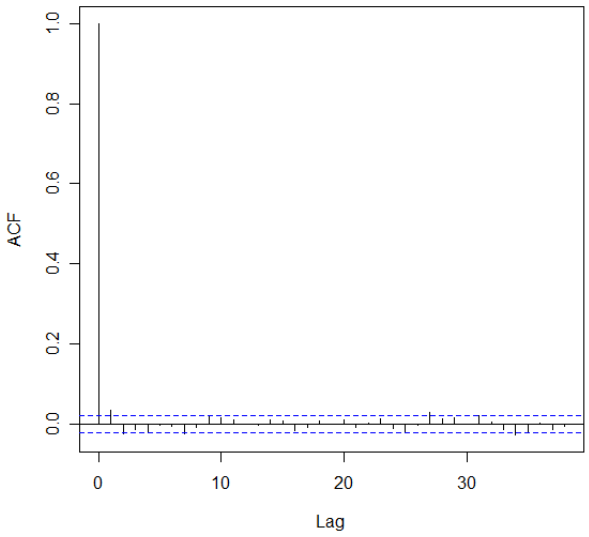

Figure 5.1: Auto-correlation Plot of Boeing Daily Returns

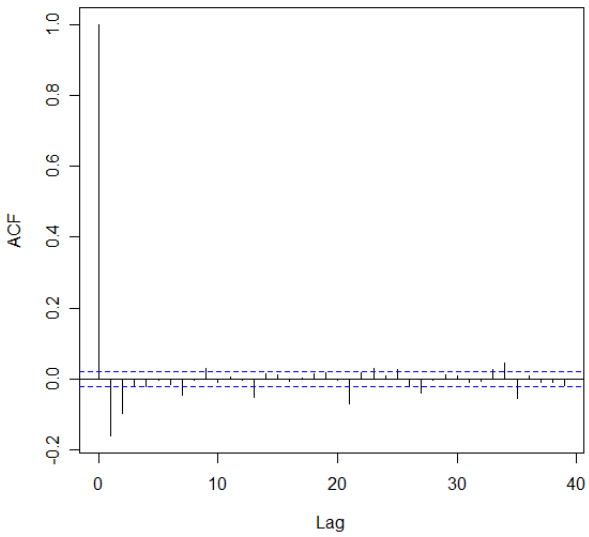

Figure 5.2: Auto-correlation Plot of Wal-Mart Daily Returns

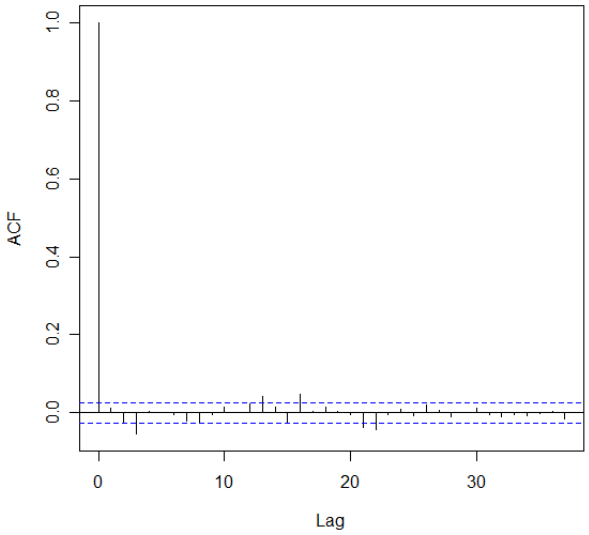

Figure 5.3: Auto-correlation Plot of Intel Daily Returns

Chapter 6

Conclusions

6.1 Results

We have seen that the log returns of stock are not fully describes by a models which strictly allow only for independent and identically distributed normal random variables. At a monthly frequency it was shown that the returns while exhibiting some signs of non normality did for the most part show the characteristics of normal distribution. As we investigated higher frequencies we discovered considerable difference to a normal distribution.

At a high frequency the returns show alot of negative skewness. They were also heavily kurtotic with pointed peaks and long tails. The indication was that they displayed more extreme values particularly as losses that the normal distribution could capture.

Taking the empirical distribution to be a mixture of GPDs in the tails and a normal distribution at its centre we applied the Peak Over Threshold

method to investigate the extremes. We found that the tails of losses could find a close fit in a GPD with a threshold chosen sufficiently high enough the satisfy the Pickands, Balkema and de Haan theorem. The corresponding MLEs of the parameters calculated and the fitted distributions were plotted successfully.

From the basic time plots of the returns there was noticable periods of high and low volatility which seem to alternate. Also when the serial correlations of the returns were calculated and plotted there was substantial values for correlation at the higher frequencies particularly for daily returns.

6.2 Conclusions

We are satisfied to say that for high frequencies the returns of the data studied was not normal enough to accept the Gaussian distribution to model it. The returns have heavy tails and so the generalised Pareto distribution is more acceptable to capture the greater and more frequent extremes that returns show. It is also true to say that the tail of return losses is heavier than the tail of return profits.

There is also evidence of correlation between daily returns at regular interval. The autocorrelations give evidence towards patterns in the data over time. These periods of heteroskedasticity merit further investigation.

Bibliography

Bensalah, Y.(November 2000), *Steps in Applying Extreme Value Theory to Finance: A Review.*, Bank of Canada Working Paper 2000-20.

Bera A., Jarque C. (1981). *Efficient tests for normality, heteroskedasticity and serial independence of regression residuals: Monte Carlo evidence.* Economics Letter 7, 313 - 318.

Carmona, R. (2004), *Statistical Analysis of Financial Data in S-Plus*, Springer.

Coles, S (2001), *An Introduction to Statistical Modelling of Extreme Values.*, Springer Series in Statistics.

D'Agostino, R. and Stephens, M. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, Inc.

Duong, T. (2001), *An introduction to kernel density estimation*, www.maths.uwa.edu.au/ duongt/seminars/intro2kde/

Embrechts, P., Kluppelberg C., and Mikosch T. (1997), *Modelling Extremal Events for Insurance and Finance.*, Springer.

Hellström, T. (1998), *A Random Walk Through the Stock Market*, Umea University,Sweden.

McNeil, A. (1996), *Estimating the Tails of Loss Severity Distributions using Extreme Value Theory.*, www.math.ethz.ch/m̃cneil/pub_list.html.

McNeil, A. (1999), *Extreme Value Theory for Risk Managers.*, www.math.ethz.ch/ mcneil/pub_list.html.

McNeil, A. and Saladin, T. (1997), *The Peak Over Thresholds Method for Estimating High Quantiles of Loss Distributions.*, www.math.ethz.ch/ mcneil/pub_list.htm

Sarma, M. (2002), *Extreme Value Theory and Financial risk management*, http://www.stat.tamu.edu/ jianhua/stat689-08sp/reference/Sarma02.pdf

Shapiro, S. and Wilk, M. (1965). *An analysis of variance test for normality (complete samples)*, Biometrika, 52, 3 and 4, 591-611.

Thode, H. (2002), *Testing for Normality*, Marcel Dekker Inc.

Tsay, R. (2002), *Analysis of Financial Time Series*, John Wiley and Sons.

Weisstein, Eric W. *Skewness.* From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Skewness.html

Zivot, E. and Wang, J. (2003), *Modelling Financial Time Series with S-Plus*, Springer.

Websites:

http://finance.yahoo.com/q?

http://www.itl.nist.gov/div898/handbook/eda/eda.htm

http://www.walmartstores.com/AboutUs/7603.aspx