ALGORITHMIC FINANCE

Christopher Ting
Lee Kong Chian School of Business
Singapore Management University

Version 1.1
April 10, 2018

CONTENTS

CHAPTER 3

STOCK PRICES AND LOG RETURNS

3.1 Introduction

In Finance, the basic unit of producers of economic goods and services is a company. Owners of a company invest their money, time and energy to produce goods and services to generate wealth. When they do not have sufficient cash or capital to invest or to expand the business, they borrow from others. There are a few options to raise the capital:

- Take a loan from the bank
- Issue bonds
- Conduct private placements of shares
- Obtain stock listing in a stock exchange to issue shares to the public

A loan is a bilateral contract between the company and the bank while a bond is a contract between the company and a number of financial institutions or retail investors. In return, company must pay interest to the bank and the bond holders. Private placement is a business deal between the company, its business partners or venture capitalists. In a private placement, company shares are sold at a fixed price after negotiation. It is an exclusive share offer. In contrast, stock listing on an exchange via initial public offering (IPO) is non-exclusive. Members of the public who want to be a co-owner of the company can bid for the shares.

Figure 3.1 A specimen of GE common stock certificate (source: NYSE).

A share of a stock is a contract that confers company ownership to shareholders under well-specified terms. Shareholders are not liable to meet the demand of the company's creditors should the company go bust. They have the right to vote during annual and extraordinary general meetings. One share is entitled to one vote and it is a slice of the company's equity, which is whatever left over after the company's liability is fully accounted for by the company's asset.

It is important to recognize from the investment standpoint that the main reason for investing in stocks is that the company is profitable in its business, and the equity remains positive and growing. Mature companies usually distribute earnings as dividend or other types of distribution such as bonus shares to the shareholders.

On the other hand, shareholders are not answerable to the company's creditors. On accounting terms, a company's equity—asset less liability—can be negative. Even if the company has more liability than asset, shareholders do not have to make up for the shortfall. So, the value of a share cannot be negative. Since the share value is always positive, the share price must also be strictly positive.

3.2 Historical Share Prices and Stock Splits

As a publicly listed company, the stock is open for trading for several hours each business day on an exchange. The last traded price of the day is typically recorded in the press. It is important to note that the last traded price does not occur exactly at the closing time of the exchange. For example, on the New York Stock Exchange, the closing time is 4:00 PM Eastern Time. Some stocks may have 4:00 PM when the last trades

occur. Other stocks may have the last trade any time before 4:00 PM. Nonetheless, the last trade price is taken as the closing stock price for the day.

Note that the time t is implicitly assumed to be progressing at a fixed quantum. If P_t is the closing price at time or day t, then P_{t-1} is the closing price a day earlier, and P_{t+1} is the closing price at day later. The day here refers to business day when trading occurs. Sundays, Saturdays, and public holidays are non-business days. If P_t is the closing price for Friday, then P_{t+1} denotes the closing price for Monday.

Though stocks were traded since the 17th century in Holland, a comprehensive and systematic archive of stock prices however, dates back to 31 December, 1925 only in CRSP's database. Take General Electric as an example. This company has an illustrious history going back to 1890. It was founded by the renowned inventor Thomas Edison. Two years later, General Electric was formed after Edison's company was merged with its rival, Thomson-Houston Electric Company. Shares were issued (see Figure 3.1) and started trading on NYSE. On its first day of trading, only 50 shares changed hands at $108 per share. In May 1896, General Electric was selected as one of the 12 original companies in the newly formed Dow Jones Industrial Average Index.

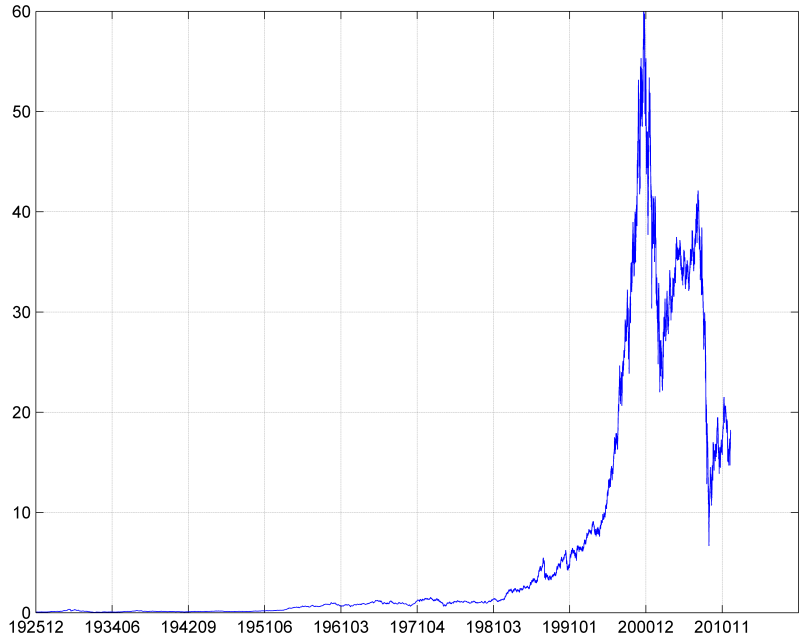


Figure 3.2 Adjusted closing prices of GE from end of 1925 through end of 2011 (data source: CRSP).

The historical closing prices of General Electric are plotted in In Figure 3.2. It is evident from the time-series plot that the stock price increases exponentially from end of December till the all-time high of $60 per share on August 28, 2000. On March 3, 2009, the share price dropped to $7.01, or more than 88% from the all-time high.

By eyeballing the data, we find that GE was actually traded at hundreds of dollars in the 1920's. But Figure 3.2 shows that the share price was less than a dollar. The

reason is that the time series of stock prices and volumes must be adjusted for stock splits. When a company's share price increases rapidly, it becomes "expensive." The company decides to split one share into x shares, thereby reducing the share price by x times. For example, GE's most recent stock split occurred on May 8, 2000 when a share split into 3 shares. Everything else being equal, the share price must therefore be 1/3 of the pre-split or "old" price, so that the dollar value of holding the shares remains unchanged. In other words, the market capitalization, which is the number of shares N_{old} times the stock price P_{old}, i.e.

$$\text{MC}_{\text{old}} = N_{\text{old}} \times P_{\text{old}}\,,$$

must not change under a stock split.

Now, the new number of new shares is $N_{\text{new}} = xN_{\text{old}}$. It follows that

$$\text{MC}_{\text{old}} = xN_{\text{old}} \times \frac{P_{\text{old}}}{x} = N_{\text{new}} \times P_{\text{new}}\,,$$

where $P_{\text{new}} = P_{\text{old}}/x$.

Suppose there were n stock splits in the past, and the split ratios were x_i, $i, = 1,2,\ldots,n$, respectively. How should the historical prices be adjusted when more than one stock split occurred? To answer this question, consider the diagram in Figure 3.3, where 3 stock splits had occurred at times t_1, t_2, and t_3, with t_3 being the most recent.

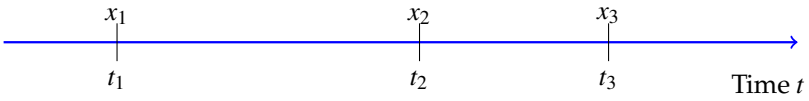


Figure 3.3 Multiple stock splits.

To compute the adjustment factor, we start from the most recent stock split at time t_3. The stock split at time t_3 requires prices to be adjusted from $t_3 - 1$ all the way back into the historical past. Similarly, the stock split at t_2 necessitates the adjustment from $t_2 - 1$ backward in dates; and the stock split at t_1 needs further adjustment to the historical prices from $t_1 - 1$ backward. In other words, from t_2 to $t_3 - 1$ (inclusive of both dates), the price adjustment factor is $d_3 = x_3$. For the time period from t_1 to $t_2 - 1$, the adjustment factor is $d_2 = x_2 d_3 = x_2 x_3$; and finally before t_1, the adjustment factor is $d_1 = x_1 d_2 = x_1 x_2 x_3$.

At and after the time t_n of the most recent stock split, the closing price needs no adjustments. For convenience, we define $d_{n+1} = 1$. The cumulative adjustment factor after i stock splits is thus given by

$$d_i = x_i d_{i+1} = x_1 x_2 \cdots x_i\,,$$

for $i = n, n-1, \ldots, 1$. Given that each stock split ratio is 2 or higher, it is obvious that

$$1 = d_{n+1} < d_n < d_{n-1} < \cdots < d_2 < d_1.$$

In other words, when many stock splits had occurred, the adjustment factor increases in a stepwise fashion backward in time. So, at the chronological beginning of the time series of share prices, the adjustment factor is the largest, and that is why before 1960, the adjusted prices of GE are less than a dollar as in Figure 3.2.

The ratios of stock splits (that are integers) for GE are listed in the table below.

Split Date	Split Ratio	Cumulative Adjustment Factor
		4,608
19260527	4	1,152
19300128	4	288
19540614	3	96
19710608	2	48
19830602	2	24
19870526	2	12
19940516	2	6
19970512	2	3
20000508	3	1

In this example, a total of 9 stock splits had occurred. Using the method described, we have $d_{10} = 1$. The most recent stock split occurred on May 8, 2000, when a share was split into 3 shares. Accordingly, $d_9 = 3$, which applies to prices from May 12, 1997 to a business day before May 8, 2000. The next most recent stock split gives rise to $d_8 = 2 \times 3 = 6$. With $d_4 = 96$, $d_3 = 288$ and so on, and since adjustment is carried out by dividing the pre-split stock price by the applicable adjustment factor, it is easy to appreciate why the adjusted prices become smaller and smaller.

3.3 Log Prices and Log Returns

In Figure 3.4, the log price $p = \ln(P)$ is plotted instead. The logarithm function transforms the exponentially increasing price P into a log price p that appears more balanced in highlighting the price fluctuation.

It is noteworthy that in contrast to the price P, the log price p can be negative, for the logarithm function yields negative values when P is less than \$1. In light of Figure 3.4, a great deal of ups and downs become visible in the early part of the time series. There was a rapid increase in the share price since December 1925 till 1929 when the peak was reached on August 19. What follows was the Great Depression in 1930's, during which the share price dropped by more than 90% from the peak.

The time series of log prices demonstrates clearly that risky asset such as GE stock produces a good return over a *long* period of time. In the worst case scenario, suppose

Figure 3.4 Adjusted logarithmic closing prices of GE from end of 1925 through end of 2011 (data source: CRSP).

an investor bought GE shares at the height of the 1920s' bubble, and sold the shares at the bottom of the 2007–2009 financial crisis, the log price difference would be

$$\ln(7.01) - \ln(398.75/1152) = 3.0083.$$

Note that 7.01 is GE's closing price on March 3, 2009 mentioned previously. The last traded price on Aug 19, 1929 is 398.75, and 1152 is the applicable price adjustment factor.

Now, the log price difference is in related to return r. To substantiate this claim, consider the return

$$r = \frac{P_u - P_t}{P_t} = \frac{P_u}{P_t} - 1.$$

In other words,

$$\frac{P_u}{P_t} = 1 + r,$$

where the time u is later than t. Apply the logarithm on both sides, we find

$$\ln(P_u) - \ln(P_t) = \ln\left(\frac{P_u}{P_t}\right) = \ln(1+r).$$

Accordingly, the log price difference is related to return r by the logarithmic function of r.

If the return r is obtained over T number of years, the annualized return r_a can be backed out by the following formula,

$$(1+r_a)^T = 1 + r = \frac{P_u}{P_t}.$$

This equation suggests that r_a is the compound annual growth rate averaged across T years. To back out r_a, we rewrite the equation as

$$\ln(1+r_a) = \frac{\ln(P_u) - \ln(P_t)}{T}.$$

It follows that

$$r_a = \exp\left(\frac{\ln(P_u) - \ln(P_t)}{T}\right) - 1.$$

In the worst case scenario described above, we have $\ln(P_u) - \ln(P_t) = 3.0083$. From Aug 19, 1929 (time t) to Mar 3, 2009 (time u), for which the number of years T is approximately 80. Inserting these data into the equation, we find that

$$r_a \approx \exp\big(3.0083/80\big) - 1 = 3.8320\%.$$

The capital appreciation of GE stock over these 80 years was 3.8320% per year.

We note that

$$\frac{3.0083}{80} = 3.7604\%,$$

which is a mere difference of 0.0716 percentage points from r_a. To account for the small difference, we perform Taylor's expansion and obtain

$$\ln(1+r_a) = r_a - \frac{1}{2}r_a^2 + \frac{1}{3}r_a^3 + \cdots$$

For small r_a, we have

$$\ln(1+r_a) \approx r_a\,.$$

Accordingly,

$$\frac{\ln(P_u) - \ln(P_t)}{T} \approx r_a\,.$$

Motivated by this finding, we proceed to define continuously compounded rate of return r_c as follows. Given two prices P_u and P_t at times u and t, which are T years apart, i.e., $u - t = T$ years. The log return r_ℓ is defined as the difference of log prices:

$$r_\ell := \ln(P_u) - \ln(P_t)\,.$$

The rate of log return, also known as the continuously compounded rate of return r_c, is defined as

$$r_c = \frac{r_\ell}{T} = \frac{\ln(P_u) - \ln(P_t)}{T}\,.$$

A few simple steps lead to

$$P_u = P_t e^{r_c T}\,. \tag{3.1}$$

This equation suggests that, on average, the stock price increases exponentially from the initial price of P_t over T years, i.e., $u - t = T$. Indeed, Figure 3.2 provides an example of the exponential growth.

3.4 Modeling Stock Price Movements

Earlier, by introducing a continuously compounded rate of return, a simple model of stock price is obtained.

$$P_t = P_0 e^{r_c t}. \tag{3.2}$$

This simple equation is a rewrite of equation (3.1), with t replaced by 0, u replaced by t, and hence $T = t - 0 = t$.

The model is crudely simple. The randomly wiggling and undulatory nature of the path taken by the stock price is missing in the model. As a matter of fact, model (3.2) is deterministic. Going forward in time, it can only increase and not decrease. Moreover, with P_t being an exponential function of time t, it is smooth; P_t can be differentiated infinitely many times, i.e. $\dfrac{d^h P_t}{dt^h} = r_c^h P_t$, for $h = 1, 2, \ldots, \infty$. Clearly, the price path in Figure 3.2 is anything but smooth.

A natural improvement to model (3.2) is to postulate that the log return is random. Specifically, we alter the constant r_c into a function of the random variable X_t

$$r_c(X_t) = \bar{r} + \sigma X_t. \tag{3.3}$$

In words, we let the log return r_c to fluctuate with respect to an "average" value $\bar{r}$. The fluctuation is captured by the random variable, X_t, and the magnitude of fluctuation is given by a constant parameter σ. It is easy to see that r_c as defined in equation (3.3) is a generalization of model (3.2). If we set σ to zero, then $r_c(X_t) = \bar{r}$ is a constant and model (3.2) is recovered. With random function (3.3), we have

$$P_t = P_0 e^{\bar{r}t + \sigma t X_t}. \tag{3.4}$$

The legacy from the deterministic model (3.2) can still be seen in $P_0 e^{\bar{r}t}$.

To gain insight into model (3.4), we partition the time interval from time 0 to time t by n subperiods. The duration of each time interval Δt is

$$\Delta t = \frac{t}{n}.$$

Altogether, there are n intervals of equal length. A pictorial illustration of the uniform partition is depicted in Figure 3.5. To simplify the notation, we write $\tau_k = k\Delta t$, where

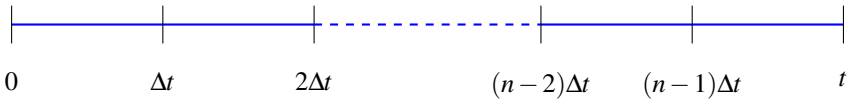


Figure 3.5 Partition of the time from 0 to t by n equal intervals.

$k = 0, 1, 2, \ldots, n-1, n$. In this notation, $\tau_0 = 0$, and $\tau_n = t$. With regard to this partition, there are n random variables X_{τ_k}, where $k = 1, 2, \ldots, n$.

Consequently, the stock price at time τ_1 according to model (3.4) is

$$P_{\tau_1} = P_{\tau_0} e^{\bar{r}\Delta t + \sigma \Delta t X_{\tau_1}}.$$

In general,

$$P_{\tau_k} = P_{\tau_{k-1}} e^{\bar{r}\Delta t + \sigma \Delta t X_{\tau_k}}. \tag{3.5}$$

By repetitive substitution, we find that

$$P_t = P_{\tau_n} = P_0 e^{\bar{r}t + \sigma \Delta t \sum_{i=1}^{n} X_{\tau_i}}.$$

So far, we have not specified the behavior of the random variable X_t. We make further assumption on X_{τ_i} as follows:

$$X_{\tau_i} := \frac{1}{\sqrt{\Delta t}} Y_{\tau_i} = \sqrt{\frac{n}{t}} Y_{\tau_i}, \qquad i = 1, 2, \ldots, n, \tag{3.6}$$

where Y_{τ_i} is a Bernoulli random variable, which takes the value of either 1 or -1 with equal probability. We also assume that Y_{τ_i} is independent of each other.

Now, the discrete Bernoulli random variable has mean 0 and variance 1, i.e., $\mathbb{E}\big(Y_{\tau_i}\big) = 0$, and $\mathbb{V}\big(Y_{\tau_i}\big) = 1$. Consequently,

$$\mathbb{E}\big(\ln(P_t) - \ln(P_0)\big) = \bar{r}t, \tag{3.7}$$

$$\mathbb{V}\big(\ln(P_t) - \ln(P_0)\big) = \sigma^2 (\Delta t)^2 \frac{n}{t} \sum_{i=1}^{n} \mathbb{V}\big(Y_{\tau_i}\big) = \sigma^2 \frac{t^2}{n^2} \frac{n}{t} n \tag{3.8}$$

$$= \sigma^2 t. \tag{3.9}$$

In other words, the expected value of the log return is $\bar{r}t$, and the variance of the log return is $\sigma^2 t$. The parameter σ^2 can be interpreted as the rate of variance.

In fact, the constant $1/\sqrt{\Delta t}$ in Equation (3.6) is deliberately included so that **the variance of log return scales linearly with time** t. The paradigm in which we operate is the random walk model. Specifically, from Equations (3.5) and (3.6), we have the random walk model as follows:

$$\ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right) = \bar{r}\Delta t + \sigma \Delta t X_{\tau_k} = \bar{r}\Delta t + \sigma \sqrt{\Delta t} Y_{\tau_k}.$$

It can be readily shown that $\mathbb{E}\left(\ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right)\right) = \bar{r}\Delta t$, and that $\mathbb{E}\left(\ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right)\right) = \sigma^2 \Delta t$. Since the variance increases linearly with time t, random walks, being a model for log prices, are non-stationary.

Now, if we set the time scale in such a way that $\Delta t = 1$, then the 1-period log return $r_{\tau_k} := \ln\left(P_{\tau_k}\right) - \ln\left(P_{\tau_{k-1}}\right)$ is a random walk with drift $\bar{r}$. In other words, the deviation from the mean $\bar{r}$ is purely random:

$$r_{\tau_k} - \bar{r} = \sigma Y_{\tau_k}. \tag{3.10}$$

3.5 Simulating Stock Price Movements and Reality Check

A simulation of the price process model (3.5) with Bernoulli fluctuation is shown in Figure 3.6. The simulated price series looks realistic and qualitatively similar to the

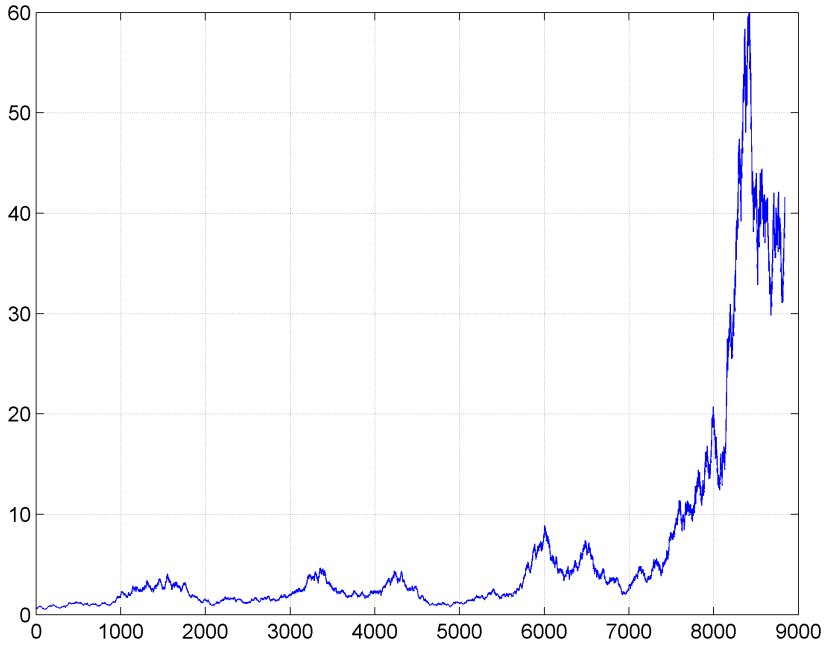


Figure 3.6 Simulated prices using model (3.5)

time series of GE prices in Figure 3.2. However, this model of price process has a fundamental flaw. Namely, as seen from equation (3.5), the log return is restricted to take two values only, either $\bar{r}\Delta t - \sigma\sqrt{\Delta t}$ or $\bar{r}\Delta t + \sigma\sqrt{\Delta t}$, since Δt, $\bar{r}$, and σ are fixed. In reality, the log return of any stock such as GE can have many different values.

Therefore, instead of the over-simplified model to drive the stock price fluctuation, we substitute the Bernoulli random variable Y_{τ_i} in Equation (3.10) by a standard normal random variable, which too has zero mean and unit variance. It turns out that the resulting price model is the discretized version of the well known geometric Brownian motion, for which the log return is a normally distributed random variable.

Definition 3.1 A discretized geometric Brownian motion is a purely random process by which the log price is purely random in such a way that the log retun is a standard normal random variable Y_t in the one-period setting:

$$Y_t \sim N(0,1).$$

■

In other words, the deviation of one-period log return from its mean, Equation (3.10), is pure noise, i.e.,

$$X_t := r_t - \bar{r} = \sigma Y_t. \tag{3.11}$$

Now, does the geometric Brownian motion correspond to reality? In Figure 3.7, the histogram of the log returns of GE is plotted. It displays the number or frequency of realized log returns with respect to the discrete interval of their values. For comparison, values of a normally distributed random variable are generated, and their histogram is superimposed as a solidly filled plot. The total number of these randomly generated values is the same as GE's total number of log returns observed over the sample period. The simulated values are generated in such a way that their mean and variance are the same as those of GE's log returns.

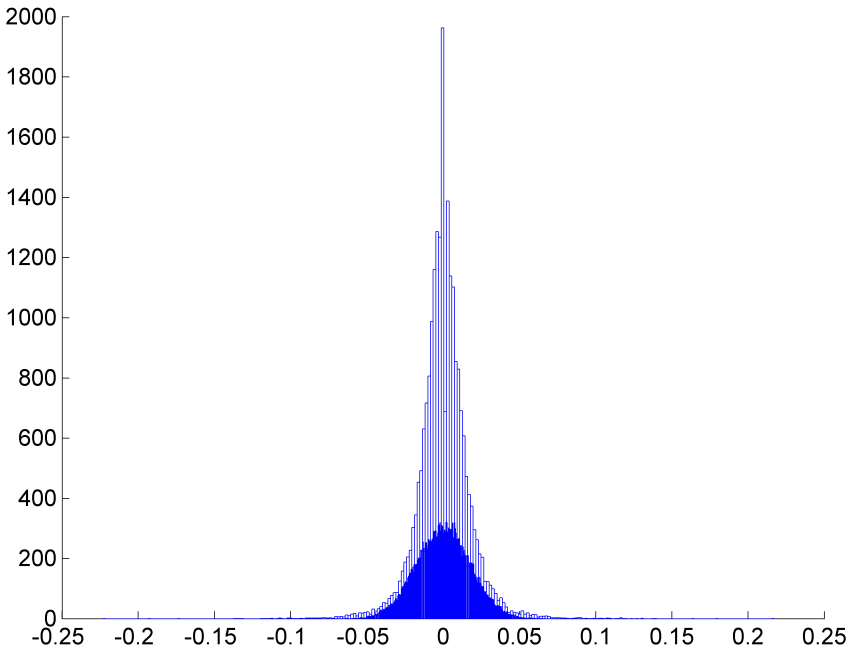


Figure 3.7 Histogram (in vertical strips) of GE's log returns and that (in solidly filled shape) of normally distributed random numbers, which are generated in such a way that their mean and variance match those of GE's log returns.

In comparison to the simulated histogram, the realized log returns have many "outliers" in the sense that there are more extreme values. For example, it is noticeable that the log returns of about -2% or lower occur more frequently than normally distributed random values do. Similar thing can be said of returns that are 2% or higher. Conversely, small returns around the mean are more frequent than normal distribution. It is intuitively evident that the distribution of GE's daily log returns is not a normal distribution. Hence, the stock price process is most likely not a geometric Brownian motion.

3.6 Statistical Tests of Normality

Jarque and Bera [JB87] provide a test to infer whether a sample of log returns is drawn from a normal distribution. Recall that a normal distribution is defined by its mean

μ and variance σ^2. Since the distribution is symmetric with respect to the mean, any higher odd-order (centralized) moment is zero. The skewness of a random variable X, which is of the third order, is defined as

$$\gamma = \frac{\mathbb{E}\big((X-\mu)^3\big)}{\sigma^3}.$$

The skewness measures the slant of the distribution. It is negative when the distribution is skewed toward the left, i.e., there are "outliers" to the left of the mean. Conversely, a positive skewness indicates the presence of extreme values to the right of the mean. Being symmetric, the skewness of the normal distribution is zero.

On the other hand, all the even-order (centralized) moments of a normally distributed random variable are not zero. In particular, the kurtosis, which is defined as

$$\kappa = \frac{\mathbb{E}\big((X-\mu)^4\big)}{\sigma^4},$$

is a fourth-order statistic, and it measures the frequency of extreme values expected of a distribution. For the normal distribution, the kurtosis is 3.

To examine whether a sample of T observations is normally distributed, we consider the Jarque-Bera statistic:

$$\mathrm{JB} = \frac{T}{6}\left(\tilde{\gamma}^2 + \frac{(\tilde{\kappa}-3)^2}{4}\right). \tag{3.12}$$

Here, $\widehat{\gamma}$ is the sample skewness and $\widehat{\kappa}$ is the sample kurtosis, which are estimated in the following way. First, the sample average is estimated:

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

Next, the estimate for the variance is based on the following estimator:

$$\tilde{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\big(x_t - \bar{x}\big)^2.$$

Finally, the estimate for the skewness is obtained as

$$\tilde{\gamma} = \frac{\frac{1}{T}\sum_{t=1}^{T}\big(x_t-\bar{x}\big)^3}{\tilde{\sigma}^3},$$

and the sample kurtosis is computed as

$$\tilde{\kappa} = \frac{\frac{1}{T}\sum_{t=1}^{T}\big(x_t-\bar{x}\big)^4}{\tilde{\sigma}^4}.$$

As can be seen from expression (3.12), a large value of $\tilde{\gamma}$ and/or a large difference of $\tilde{\kappa}$ from 3 will lead to a large value for the Jarque-Bera statistic. Since the skewness

	observations	skewness	kurtosis	Jarque-Bera statistics
daily	22,776	0.025999	13.13398	97,462.5
weekly	4,486	-0.04181	10.00023	9,160.8
monthly	1,032	-0.31251	7.77295	996.4
quarterly	344	-0.27640	6.66580	197.0
biannual	172	-0.94224	6.47470	112.0
yearly	86	-0.88798	4.26247	17.0

Table 3.1 Results of Jarque-Bera tests for GE's log returns at different frequencies.

and kurtosis are, respectively, 0 and 3 for the normal distribution, a large Jarque-Bera statistic provides a measure for the deviation from normality.

To conduct the Jarque-Bera test, we set the null hypothesis as $H_0 : \text{JB} = 0$. The alternative hypothesis is $H_1 : \text{JB} \neq 0$. Jarque and Bera show that the JB statistic is a χ^2_2 distributed random variable with 2 degrees of freedom [JB87]. We perform 6 separate tests for daily, weekly, monthly, quarterly, biannual, and yearly log returns. Table 3.1 shows the relevant statistics in the context of Jarque-Bera tests.

The critical or cut-off value of chi-square statistic at the 0.5% significance level is 10.597. Since all the Jarque-Bera statistics are greater than 10.597, there is evidence to reject the null hypothesis of normality. It is also noteworthy that the kurtosis decreases monotonically as the sample frequency increases.

3.7 Autocorrelation of Log Returns

The mean, variance, skewness, and kurtosis, do not take the temporal structure of the log returns into account. The time t of the log return r_t is used purely as the index in the summation when these descriptive statistics are computed. The histogram, too, does not provide information about the temporal sequence of the log return.

This section provides a different statistical tool to ferret out any insightful information that might be hidden in the temporal realm of r_t.

Definition 3.2 We define the autocorrelation of a time series x_t as the correlation of x_s with x_t, for all s and t.

$$\rho(s,t) := \frac{\mathbb{C}(x_s, x_t)}{\sqrt{\mathbb{V}(x_s)}\sqrt{\mathbb{V}(x_t)}}.$$

■

It is evident that $\rho(t,t) = 1$. The question of interest is $\rho(s,t)$ for $s = t-1, t-2, \ldots, t-k$. Accordingly, we have the following definition.

Definition 3.3 We define an autocorrelation function (ACF) up to lag k as follows:

$$\text{ACF}(h) := \frac{\mathbb{C}\big(x_{t-h}, x_t\big)}{\sqrt{\mathbb{V}\big(x_{t-h}\big)}\sqrt{\mathbb{V}\big(x_t\big)}}, \qquad \text{for } h = 0, 1, 2, \ldots, k.$$

■

To simplify the analysis, an important assumption of homoskedasticity is made. Namely, for all h,

$$\mathbb{V}\big(x_t\big) = \mathbb{V}\big(x_{t-h}\big) = \sigma^2. \tag{3.13}$$

Under this assumption, the autocorrelation function is written as

$$\text{ACF}(h) = \frac{\mathbb{C}\big(x_{t-h}, x_t\big)}{\sigma^2}, \qquad \text{for } h = 0, 1, 2, \ldots, k.$$

Following [BJR94], given a time series of T observations $\{x_t\}_{t=1}^{T}$, the sample estimate of ACF(h) for $h = 0, 1, 2, \ldots, k$, can be obtained as

$$\gamma_h = \frac{c_h}{c_0},$$

where

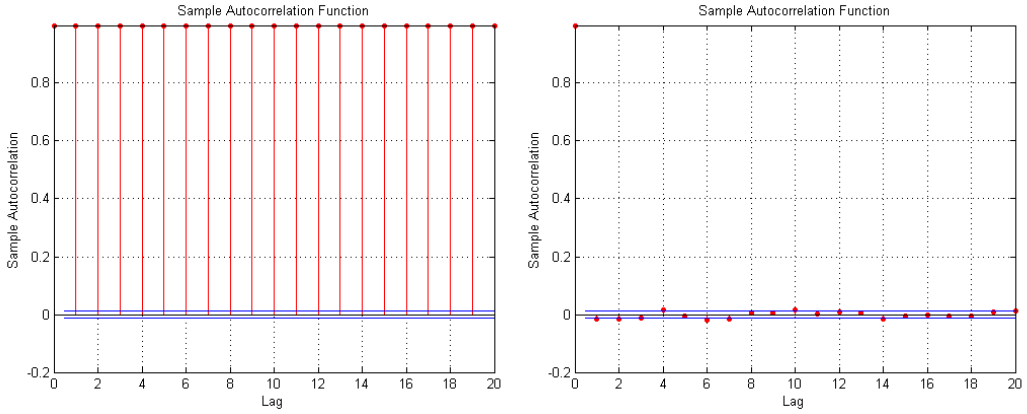
$$c_h = \frac{1}{T-h} \sum_{t=h+1}^{T} (x_t - \bar{x})(x_{t-h} - \bar{x}), \tag{3.14}$$

and $\bar{x}$ is the sample mean $\bar{x} = \sum_{t=1}^{T} \frac{x_t}{T}$. In Equation (3.14), note that the summation index starts from $t = h+1$. This is simply because the time series starts from $t = 1$ and x_{t-h} is meaningless if $t < h+1$.

Intuitively, γ_h for a given h is the correlation between the random variable at time t with the same random variable at a different time $t-h$. Suppose γ_h is 0.7 and the autocorrelations at other lags are all zero. Then roughly speaking, there is a 70% chance that x_t is positive in this (hypothetical) example.

Panel A of Figure 3.8 plots the sample ACF(20) of the daily log price of GE. For lag 1 to lag 20, the values of γ_h are close to 1, which provides little information about the temporal structure of p_t. A possible explanation is that the sign of log price at time $t-i$ and the sign at time t are almost always the same when i is a small number. More importantly, the value of the log price p_t at time t is usually not much different from p_{t-j} when normalized by c_0. Even despite the fact that the log price of GE can be either positive or negative as evident in Figure 3.4, the temporal structure nonetheless is such that they are highly correlated. This characteristic of $\gamma_j \approx 1$ for $j = 1, 2, \ldots$ is typical of a non-stationary time series. Although not statistically rigorous, the sample ACF does provide a quick diagnosis of whether a time series is non-stationary.

By contrast, Panel B shows that the daily log returns at different times have practically no correlations at all. Some of the γ_h are however statistically significant. As



(A) Log Price (B) Log Return

Figure 3.8 Two sample ACF(20) of GE's daily log price (left) and log returns (right).

shown by [Bar46], the variance of γ_h is well approximated by $1/T$. Consequently, a large T gives rise to a small variance, and thus it is easy for even a small γ_h estimates to exceed the two-tail critical values, which are indicated by two blue lines parallel to the horizontal axis.

Since the log return is the first difference of the log price, i.e.,

$$r_t = \Delta p_t = p_t - p_{t-1} = \ln(P_t) - \ln(P_{t-1}),$$

it can be said that the log price difference Δp_t at time t has no memory of the past log price difference Δp_{t-i}. In this context, using the past daily return to forecast the future daily return is quite futile.

3.8 Variance Ratio Test

When the daily log return r_t is treated as a random variable, the variance of a sum of q daily log returns *in sequel* is

$$\mathbb{V}\left(\sum_{t=1}^{q} r_t\right) = \sum_{t=1}^{q} \mathbb{V}(r_t) + 2\sum_{t=1}^{q}\sum_{s<t}^{q} \mathbb{C}(r_s, r_t).$$

To simplify the analysis, two assumptions are made, as we did before:

1. Zero covariance: $\mathbb{C}(r_s, r_t) = 0$ for any $s \neq t$

2. Homoskedasticity: $\mathbb{V}(r_t) = \sigma^2$

Definition 3.4 The non-overlapping q-daily log return is defined as

$$\begin{aligned} r_t(q) := \sum_{j=1}^{q} r_{t-q+j} &= \big(\ln P_t - \ln P_{t-1}\big) + \big(\ln P_{t-1} - \ln P_{t-2}\big) + \cdots \\ &\quad \cdots + \big(\ln P_{t-q+2} - \ln P_{t-q+1}\big) + \big(\ln P_{t-q+1} - \ln P_{t-q}\big) \\ &= \ln P_t - \ln P_{t-q}. \end{aligned}$$

Definition 3.5 The variance ratio for the q-period log return is defined as

$$\mathrm{VR}(q) = \frac{\mathbb{V}\big(r_t(q)\big)}{q\sigma^2}.$$

Under the assumptions of zero covariance and homoskedasticity,

$$\mathbb{V}\big(r_t(q)\big) = \mathbb{V}\left(\sum_{t=1}^{q} r_t\right) = \sum_{t=1}^{1} \mathbb{V}(r_t) = q\sigma^2.$$

In this expression, the constant σ^2 is the variance of daily log return. It follows that $\mathrm{VR}(q)$ should be equal to one when the conditions of log returns being serially uncorrelated and homoskedastic are satisfied. The variance ratio test is a test of

$$H_0 : \mathrm{VR}(q) - 1 = 0 \qquad \text{versus} \qquad H_1 : \mathrm{VR}(q) - 1 \neq 0.$$

If the null hypothesis cannot be rejected, then it means that the two assumptions made are consistent with the reality. Conversely, a rejection of H_0 implies that either one or both of the assumptions is/are inconsistent with the data.

To set up the framework for inference, we recall a few definitions and facts. The sample mean of daily log returns is estimated as usual,

$$\widehat{r}_1 = \frac{1}{T}\sum_{t=1}^{T} r_t.$$

The sample variance of daily log returns is estimated as

$$\widehat{\sigma}_1^2 = \frac{1}{T}\sum_{t=1}^{T} \big(r_t - \widehat{r}_1\big)^2.$$

The subscript of 1 in $\widehat{r}_1$ and $\widehat{\sigma}_1^2$ is meant to indicate that these estimates are for daily log returns. By the law of large numbers, as $T \longrightarrow \infty$,

$$\mathbb{E}\big(\widehat{\sigma}_1^2\big) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big(\big(r_t - \widehat{r}_1\big)^2\Big) \longrightarrow \sigma^2.$$

This result is a reflection of the fact that the variance estimator $\widehat{\sigma}_1^2$ is a consistent estimator.

Proposition 3.6 Asymptotically, when $T \to \infty$,

$$\mathbb{V}(\widehat{\sigma}_1^2) = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\Big((r_t - \widehat{r}_1)^2\Big) \longrightarrow \frac{2\sigma^4}{T}.$$

■

Proof: From Equation (3.11), it is clear that, for the log return, the deviation from the mean, i.e., $r_t - \widehat{r}_1$, is σY_t, where Y_t is a standard normal random variable. Hence, given the assumption of zero covariance and homoskedasticity,

$$\begin{aligned}\mathbb{V}(\widehat{\sigma}_1^2) &= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\Big((r_t - \widehat{r}_1)^2\Big) = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\Big((\sigma Y_t)^2\Big)\\ &= \frac{1}{T}\mathbb{V}(\sigma^2 Y_t^2) = \frac{\sigma^4}{T}\mathbb{V}(Y_t^2)\end{aligned}$$

Now, it is a known fact in probability & statistics that if Y_t is s standard normal random variable, then Y_t^2 is a chi-square random variable with one degree of freedom. Also, the variance $\mathbb{V}(Y_t^2)$ of the chi-square random variable with one degree of freedom is 2. ■

By the central limit theorem, as $T \longrightarrow \infty$,

$$\sqrt{T}(\widehat{\sigma}_1^2 - \sigma^2) \sim N(0, 2\sigma^4).$$

Now the non-overlapping q-daily return denoted by $r_q(j)$ can be written as

$$r_q(j) = \ln P_{qj} - \ln P_{q(j-1)},$$

for $j = 1, 2, \ldots, M$, where M is the maximum number of non-overlapping q-daily returns that are obtainable from $T+1$ prices starting from P_0. The sample average of $r_{qj}(q)$ is simply q times of $\widehat{r}_1$, i.e., $q\widehat{r}_1$, in accordance to the linear scaling law, Equation (3.7). The sample variance is then estimated by

$$\widehat{\sigma}_q^2 = \frac{1}{M}\sum_{j=1}^{M}\big(r_q(j) - q\widehat{r}_1\big)^2.$$

Proposition 3.7 Asymptotically, as $M \to \infty$,

$$\mathbb{E}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{1}{Mq}\sum_{j=1}^{M}\mathbb{E}\Big(\big(r_q(j) - q\widehat{r}_1\big)^2\Big) \longrightarrow \sigma^2$$

■

Proof: The q-daily log return $r_q(j)$ for each j is a sum of q daily returns.

$$r_q(j) = r_{qj} + r_{qj-1} + \cdots + r_{q(j-1)+1}.$$

Applying Equation (3.11), we obtain

$$\begin{aligned} r_q(j) - q\widehat{r}_1 &= (r_{qj} - \widehat{r}_1) + (r_{qj-1} - \widehat{r}_1) + \cdots + (r_{q(j-1)+1} - \widehat{r}_1) \\ &= u_{qj} + u_{qj-1} + \cdots + u_{q(j-1)-1}. \end{aligned} \tag{3.15}$$

Since all the Y_t has zero covariance with each other, and since $\mathbb{E}(Y_t) = 0$,

$$\mathbb{E}\Big(\big(r_q(j) - q\widehat{r}_1\big)^2\Big) = \mathbb{E}\left(\sum_{i=0}^{q-1} u_{qj-i}^2\right) = \sum_{i=1}^{q-1} \widehat{\sigma}_1^2 = q\widehat{\sigma}_1^2(j).$$

The index j indicates that all the dispersions Y_t belong to the j-th sample. Consequently,

$$\mathbb{E}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{1}{M}\sum_{j=1}^{M} \widehat{\sigma}_1^2(j).$$

Therefore, when M is large, the expected value of q-daily variance divided by q approaches the true value of the daily variance σ^2. ∎

Proposition 3.8 Since $Mq = T$, the asymptotic limit of the variance of $\widehat{\sigma}_q^2$ is

$$\mathbb{V}\left(\frac{\widehat{\sigma}_q^2}{q}\right) = \frac{1}{M^2q^2}\mathbb{V}\left(\sum_{j=1}^{M}\big(r_q(j) - q\widehat{r}_1\big)^2\right) \longrightarrow \frac{2q\sigma^4}{T}.$$

∎

Proof: In view of Equation (3.15), we have M chi-square random variables, and the degrees of freedom of each is q.

$$\begin{aligned} \frac{1}{M^2q^2}\mathbb{V}\left(\sum_{j=1}^{M}\big(r_q(j) - q\widehat{r}_1\big)^2\right) &= \frac{1}{Mq^2}\mathbb{V}\left(\sum_{j=1}^{q} Y_j^2\right) = \frac{1}{(Mq)q}\mathbb{V}\left(\sum_{j=1}^{q} Y_j^2\right) \\ &= \frac{1}{Tq}\mathbb{V}\left(q\widehat{\sigma}_1^2 Y_j^2\right) = \frac{q}{T}\widehat{\sigma}^4\mathbb{V}\left(Y_j^2\right) \\ &= \frac{2q\widehat{\sigma}^4}{T}. \end{aligned}$$

∎

By the central limit theorem, as $M \longrightarrow \infty$,

$$\sqrt{Mq}\left(\frac{\widehat{\sigma}_q^2}{q} - \sigma^2\right) \sim N\big(0, 2q\sigma^4\big).$$

To perform the test, we define the sample statistics

$$J_d(q) := \frac{\widehat{\sigma}_q^2}{q} - \widehat{\sigma}_1^2;$$
$$J_r(q) := \frac{\widehat{\sigma}_q^2}{q\widehat{\sigma}_1^2} - 1 = \widehat{\mathrm{VR}}(q) - 1.$$

It turns out that the asymptotic distributions of $\sqrt{Mq}J_d(q)$ and $\sqrt{Mq}J_r(q)$ are normal with mean 0 and variances of, respectively, $2(q-1)\sigma^4$ and $2(q-1)$ (see Chapter 2 in [CLM97]) :

$$\sqrt{Mq}J_d(q) \sim N\big(0, 2(q-1)\sigma^4\big);$$
$$\sqrt{Mq}J_r(q) \sim N\big(0, 2(q-1)\big). \tag{3.16}$$

In light of (3.16), for $q > 1$, the z score is computed as

$$Z_q = \sqrt{qM}\frac{J_r(q)}{\sqrt{2(q-1)}} \sim N(0,1).$$

We use the daily log returns of GE to conduct the variance ratio tests[1] for $q = 2, 3, \ldots, 10$. In other words, the sample mean $\widehat{r}_1$ and sample variance $\widehat{\sigma}_1^2$ are estimated with daily log returns. In Table 3.2, we present the results of the variance ratio tests. For reference, we also tabulate the autocorrelations at first lag.

q	1	2	3	4	5	6	7	8	9	10
Obs	22,776	11,388	7,592	5,694	4,555	3,796	3,253	2,847	2,530	2,277
γ_1	-0.017	-0.048	-0.021	-0.029	-0.003	-0.037	-0.037	-0.004	0.027	-0.010
$\widehat{\mathrm{VR}}(q)$	1	1.002	0.946	0.939	0.916	0.926	0.968	0.933	0.871	0.920
Z_q	—	0.20	-4.08	-3.74	-4.46	-3.53	-1.40	-2.69	-4.85	-2.86

Table 3.2 Results of variance ratio tests based on GE's daily log returns.

We find that the variance ratios are generally above 0.9 with the exception of $q = 9$. It is clear, however, that the null hypothesis must be rejected for all q except for $q = 2$ and $q = 7$. A rejection of the null hypothesis means that either serial correlation or homoskedasticity, or both are not compatible with the empirical evidence. Though only γ_1 for each q is tabulated, it is a proxy for the order of magnitude of the other ACF(20)'s lags. The implications of these findings is that the homoskedasticity assumption as stated in equation (3.13) is likely to be the main source that causes the rejection.

[1] A reference for this part is [Lim11].

3.9 Hetroskedastic Time Series of Log Returns

When the assumption of homoskedasticity fails to hold, the time series is said to be heteroskedastic. As shown in Figure 3.9, the time series of GE's log returns exhibits non-uniform magnitude of fluctuation. Notably during early 1930's, early 2000's, and also from 2008 to 2009, the magnitude of fluctuation is a lot larger. Though less pronounced, pockets of high volatility, which is the intuitively the amplitude of log return, are still observable against the backdrop of much milder fluctuation. This temporal structure of volatility clustering is an ubiquitous feature of many financial time series.

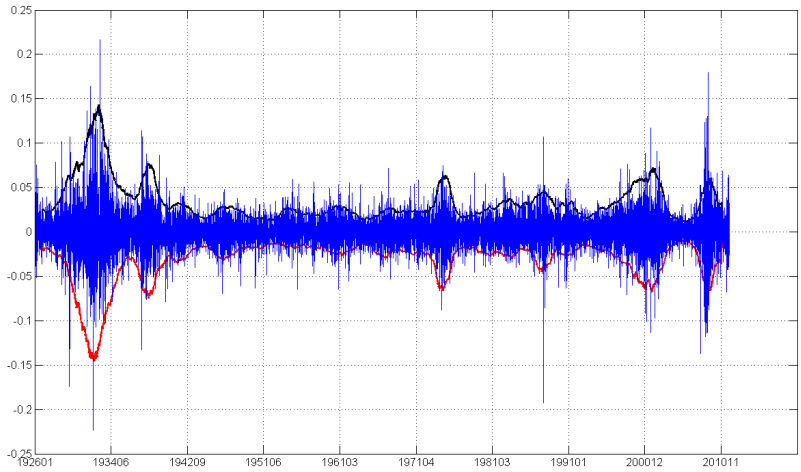


Figure 3.9 Log returns of GE from the beginning of 1926 through end of 2011, along with the upper (in black) and lower (in red) outlines of the volatility clusters.

Shown in Figure 3.9 are the outlines of the clusters. These outlines are obtained by averaging the daily log returns. The following smoothing algorithm is used:

1. Select the half window size, which is denoted by w
2. Compute the smoothed log return $\tilde{r}_t$ for each time t by

$$\tilde{r}_t = \frac{1}{n} \sum_{j=-w}^{w} r_{t+j} 1_{r_{t+j}>0} 1_{r_{t+j}<\kappa},$$

   where n is the number of daily log returns in the time window for which $-w \leq j \leq w$, and satisfy the two conditions of $r_{t+j} > 0$ and $r_{t+j} < \kappa$. The threshold parameter κ is a positive number.

3. Nonlinearly scale the smoothed log return $\tilde{r}_t$ to obtain the upper cluster outline or envelope at time t:

$$r_t^\sharp = \exp\left(\lambda \tilde{r}_t\right) \times \tilde{r}_t.$$

The amplification parameter λ is also a positive constant.

4. Compute another smoothed log return $\tilde{r}_t$ for each time t by

$$\tilde{r}_t = \frac{1}{n} \sum_{j=-w}^{w} r_{t+j} 1_{r_{t+j}<0} 1_{r_{t+j}>-\kappa}.$$

5. Nonlinearly scale the smoothed log return $\tilde{r}_t$ to obtain the lower cluster outline or envelope at time t:

$$r_t^{\flat} = \exp\left(-\lambda \tilde{r}_t\right) \times \tilde{r}_t.$$

In Figure 3.9, we use the half window size $w = 252$, which is about a calendar year. Therefore, for each day t, the smoothed log return is centered with respect to a year of log returns in the past, and a year of log returns in the future. The threshold parameter κ is set equal to 0.05 to filter out extreme log returns. To sharpen the contrast between peaks and troughs, the parameter λ is set to a value of 80.

By counting the number of peaks in the upper and lower outlines, we find 13 volatility clusters in Figure 3.9. So, over 86 years, the volatility cluster takes about 7.17 years to complete a cycle on average. What is the implication of this finding? First, multi-year oscillations suggest that volatility is cyclical in nature and it is important to know at which phase the volatility is in, whether it is moving up toward the peak, or coming down into the valley. Second, since each cycle or cluster has different length and overall amplitude of fluctuation, volatility is stochastic even at the multi-year scale.

3.10 Summary

Using GE stock as the example, this chapter provides an account of how the time series of stock prices is to be adjusted for stock splits and stock dividends. A takeaway is that it is more informative for long-term investors of GE Stock to look at the time series at the log scale, i.e., log prices.

By examining the log returns based on the variance ratio test and the Jarque-Bera statistic, we find that the log returns are by no means normally distributed and the log price is not a random walk. The implication is that there might be some pockets of opportunities for the pundits who think they have good trading strategies to "beat the market."

Using the simple autocorrelation analysis, we also show that log prices are non-stationary and log returns have virtually no serial correlation. In other words, it is very hard to beat the market, for otherwise, *too* many traders would profit from their "technical analyses," the very notion of which is self-contradictory.

Finally, this chapter also provides a simple and intuitive tool to evaluate the volatility on a macro scale. The upshot is that for the sample period from the beginning of 1926 to the end of 2011, clusters are evident and surely it is crucial to know the phase at which the volatility is at.

EXERCISES

3.1 Suppose there are 12 daily log returns, $r_1, r_2, \ldots, r_{12}$, and their values are 0.5%, 1.0%, -1.1%, -1.2%, $1.3\%, 0.7\%, -0.1\%, -0.4\%, 0.9\%, 0.6\%$, -1.5%, and -0.8%.

(a) Suppose initially the price is $P_0 = \$10$. What is the price at time 12?

(b) What is the (arithmetic) average daily return?

(c) What is the sample variance $\widehat{\sigma}_2^2$ of the bi-daily log return?

(d) What is the first-lag autocorrelation of bi-daily log returns?

(e) What is the variance ratio $\widehat{\mathrm{VR}}(3)$ of 3-daily return?

(f) What is the Z_3 score of the variance ration for 3-daily return?

(g) What are the three 4-daily log returns?

3.2 Martingale is a concept that says that given all the past prices, the prediction of tomorrow's price is the price today. Mathematically, suppose $\{P_t\}_{t=1}^T$ is a **stochastic process**. It is said to be a **martingale** if

$$\mathbb{E}(P_{t+1}|P_t, P_{t-1}, \ldots) = P_t.$$

Equivalently, since P_t is a known constant at time t,

$$\mathbb{E}\left(P_{t+1} - P_t|P_t, P_{t-1}, \ldots\right) = 0.$$

Now, consider instead the mean-squared error forecast X_t, which is expressed as

$$\mathbb{E}\left((X_t - P_{t+1})^2\middle|P_t, P_{t-1}, \ldots\right) =: f\left(P_{t+1}, X_t; P_t, P_{t-1}, \ldots\right)$$

Show that when $X_t = P_t$, the function $f\left(P_{t+1}, X_t; P_t, P_{t-1}, \ldots\right)$ is at its minimum. Specifically,

$$f\left(P_{t+1}, X_t = P_t; P_t, P_{t-1}, \ldots\right) = \mathbb{E}\left(P_{t+1}^2 - P_t^2\middle|P_t, P_{t-1}, \ldots\right).$$

3.3 Consider a drift-less random walk on a discrete grid of 18 points labeled as 0 through 17. Suppose you have equal probability of stepping up or down. The random walk will stop when you reach the boundary of 0 or 17. If you start at point 7, what is the probability that you arrive at 17 before you arrive at 0?

REFERENCES

[Bar46] M. S. Bartlett, *On the theoretical specifications of sampling properties of autocorrelated time series*, Supplement to the Journal of the Royal Statistical Society **8** (1946), 27–41.

[BJR94] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel., *Time series analysis: Forecasting and control*, 3rd ed., Prentice-Hall,, 1994.

[CLM97] John Y. Campbell, Andrew W. Lo Lo, and A. C. MacKinlay, *The econometrics of financial markets*, Princeton University Press, 1997.

[JB87] Carlos M. Jarque and Anil K. Bera, *A test for normality of observations and regression residuals*, International Statistical Review **55** (1987), 163–172.

[Lim11] Kian-Guan Lim, *Financial valuation and econometrics*, World Scientific Publising, 2011.